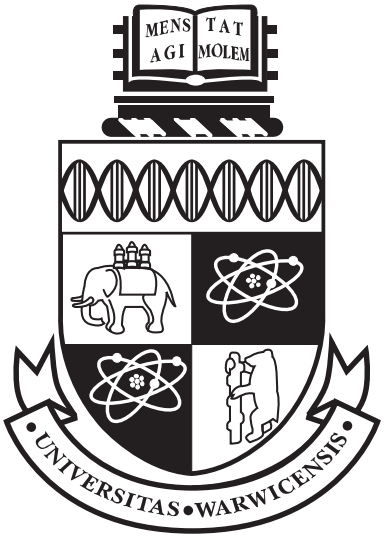

Efficient and context-dependent Bayesian model selection

by

Nick Underhill

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

June 2016

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

This thesis would not have been possible without the support and encouragement of my supervisor Jim Smith. Jim's sustained enthusiasm, despite my frequent periods of time away from the project, has enabled me to juggle the demands of my job with those of the research. Most of all, his perspective and insights on both theoretical and applied statistics, together with his willingness to take seriously my more outlandish suggestions, have made my learning experience all the more worthwhile.

I have also benefited greatly from the training given by the Academy for PhD Training in Statistics (APTS) and from discussions with members of the Statistics Department at Warwick.

I would like to thank the creators and contributors to the R programming language which I have used extensively to test ideas and to undertake simulations.

Final thanks must go to my family. Tim - your proof reading of the final draft was particularly helpful. Janet - thank you for encouraging me to start and continue with the PhD, and for putting up with my regular disappearances to work on the topic. And Henry - thanks for ensuring a healthy balance at the weekend between work on the thesis and the far more important business of remote control car racing: next challenge accepted!

This thesis was typeset with LaTeX 2_ε[1] by the author.

[1] LaTeX 2_ε is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The style package *warwickthesis* was used.

Declarations

I hereby declare that this thesis is based on my own research, except when stated otherwise. This thesis has not been submitted for a degree at another university.

Some of this work has been published, accepted for publication or submitted for publication as follows:

Some of the material in Chapters 2, 4 and 6 has been published in *Bayesian Analysis* under the title *Context-dependent score based information criteria* (Underhill and Smith [2016]). The paper was co-authored with Jim Smith but all the work is mine. A version of this material was also published as CRiSM Working Paper 14-23.

The material in Chapter 5 and some of Chapter 6 is included in a paper entitled *Modular and structure specific model selection* which is currently in preparation.

Abstract

In this thesis, we argue that the development of a number of context-dependent modifications to standard model selection approaches are warranted from an applied statistical standpoint, where we would generally accept that not only is no candidate model likely to be correct, but also that different models may be preferred for different purposes.

To achieve this we propose three types of modification. First, we consider modifications to Bayes factor selection which proceed by specialising the Bayes factor to particular variables of interest, or as an alternative, by placing vague, adaptive priors on variables of less interest.

We suggest that, particularly when the analyst wishes to assess models in light of a specific utility, scoring rules have an important role to play, and propose a new bias corrected score based information criterion which can be tailored to the utility at hand.

Finally, we present results on a modular assessment framework for 'big' models whose components can be expressed in terms of exponential families. Such an approach allows components of the broader model to be assessed individually, and the assessments combined into an overall model score. We believe that this enables the analyst to allow certain judgements about data assessment periods and exchangeability of future data to be accommodated.

We conclude with a discussion of areas for further research.

Chapter 1

Introduction

1.1 Why is context important when selecting models?

Model selection is a central concern of both theoretical and applied statistics. It is often theoretically motivated in terms of choosing the 'correct' model which the analyst believes provides a full description of the underlying generating process for the data under study. In practice, however, it is regularly undertaken in order to make sound predictions of future values of one or more specific quantities of interest. The analyst may be concerned with predicting certain marginals, conditional relationships or other quantities (for example, quantiles), but have relatively little interest in the overall performance of models across the full joint distribution. Here, the goal may be simply to pick the model which she believes will perform 'best' out of those available.

In this situation, 'best' will depend on the purpose to which the model will be put. It will depend, among other things, on the type of decision which the model informs (for example, whether the analyst wishes to provide a partial explanation of observed data, or whether she intends to use the model to make predictions of data yet unobserved) and on the utility of the decisions being made (for example, the impact of forecast errors on some variables may be more severe than on others).

In this thesis we argue that the development of *context dependent* approaches has an

important role to play, particularly from an applied statistical standpoint, where we would generally accept that not only is no candidate model likely to be correct, but also that different models may be preferred for different purposes.

1.2 Perspectives on model selection

It is well known that the fact that models being evaluated typically capture the data generating process only in an approximate way presents challenges to the procedure for model selection.

The status of these models, in terms of their claim to represent a 'true' underlying data process, may change, depending on application. Bernardo and Smith [1994] consider three possible perspectives on the models under investigation which, in turn, suggest different frameworks for the evaluation of models.

Firstly, in the *M-closed* perspective, we may have reason to believe that the set of candidate models contains the true model. This might be the assumption in a classical hypothesis testing procedure where we wish to select one of two model hypotheses which, we believe, exhaust the possibilities, or, in a Bayesian setting, to which we are prepared to assign some non zero prior probability of being the true model. Many standard model selection approaches, for example those using Bayes factors (see Kass and Raftery [1995]), assume the underlying *M-closed* context. In all but the simplest cases, however, this assumption is dubious.

A second, *M-completed*, perspective, considers the models as approximations to some *known* model where, perhaps, such approximations are being made in order that subsequent inference or analysis becomes more tractable.

In practice, we rarely operate in either of these two worlds: more usually, we operate in an *M-open* situation, in which the class of fitted models is simply a convenient proxy for an unknown true model. Here, once we accept that the model class under consideration cannot usually be guaranteed to correspond in its entirety to a comprehensive and exact representation of the modeller's belief about what might unfold, the focus immediately

shifts to identifying *which aspects* of the model performance are most important to the end user.

Rather than seeking to provide a complete and faithful portrayal or explanation of the underlying physical, causative processes, the analyst aims at releasing to her clients a model which is 'the best available and good enough' ('requisite' to use the terminology of Phillips [1982]) to enable predictions to be made within acceptable bounds, where acceptability is defined with reference to the end user's utility function.

With growing access to larger datasets, and the associated interest in 'big data' and 'big models', we argue that achievement of these more pragmatic goals will become more important. One reason for this is the appetite to exploit large datasets in the absence of a clearly defined structure *a priori*. For example, unsupervised learning (Gharamani [2004], Hastie et al. [2009], Murphy [2012]) on a large but incomplete data set provides opportunities to discover new predictive covariates, but also presents challenges in that the search for the best model is likely to be pragmatically, rather than theoretically motivated. Even in those cases where the modeller has access to greater insight into the underlying data structure and relationships, the larger the size and nature of the data set, the less feasible it is to build faithful probability structures over every aspect of the joint distribution.

1.3 Contexts for model selection

In our work within energy and financial market risk management, we have encountered a number of situations in which standard model selection procedures need to be adapted to reflect the context of the decisions being made. In this thesis, we examine three distinct contexts we frequently encounter in practice, where it is desirable to introduce alternative selection methodologies. We now outline these contexts.

1.3.1 Focus on some marginals of the joint distribution

Many standard model selection approaches, for example, Bayes factors, are implicitly concerned with the performance of the model across the full joint distribution rather than the subset of interest. In many applications, however, the consequences of poor forecasts

of some variables are often relatively minor compared to other variables.

For example, prices in the UK gas markets are highly dependent on a complex set of relationships including short term demand, commodity prices in other countries from which gas can be imported, global oil prices, and long term storage, supply and demand considerations. Cartea and Williams [2008] and Asche et al. [2006] provide good discussions of the interplay between short term and long term factors, although modelling these relationships has, arguably, recently become even more problematic due to the impact of shale gas (see, for example, Asche et al. [2012]).

In this situation, a natural approach to accommodate the variety of factors, while recognising the researcher's own uncertainty over the nature of the relationships, is to consider a high dimensional network model, which may have been built to exploit a number of promising covariates or to allow the researcher to build up a plausible set of connections between components hierarchically. See, for example, Abramson and Finizza [1991] and Yu et al. [2008] for work in this direction.

An analyst may ultimately be interested only in the dependence of price V on UK demand, D. However, rather than modelling this directly (which might require the elicitation of complex priors and likelihood specifications beyond her experience, or that of subject matter experts) she considers that a more encompassing network model is appropriate. This is because it allows her to build up a series of intermediate dependencies which are more readily modelled, understood and elicited, with the hope that in doing this, the relationship of interest can be established more accurately than through a more direct approach.

Application of Bayes factors to the resulting models will, as we comment in Chapter 3, assess the models on their overall performance expressed in terms of the full joint distribution across all variables modelled. Here, however, the analyst is interested in choosing between models based on their performance in predicting price based on a change in UK demand. In this situation, we argue that the selection criteria used should specifically focus on the relationships of interests and the utility of the model in this regard, as opposed to predictive performance on variables of less interest.

1.3.2 Utility of decision

In a number of situations, the researcher may be able to quantify the loss experienced when using a poor model, for example, through forecast error. Sometimes, this may be explicitly stated in terms of minimum regulatory requirements for model performance which must be met (see Wong [2010]), or may be implicit from the consequences of decisions made using model predictions. Here, it is clearly desirable that the selection approach is sensitive to the expected future utility resulting from use of the chosen model.

Research in this area has tended to concentrate on considering the incorporation of different utilities within the M-closed framework (for example, Bernardo and Smith [1994] for general utilities and San Martini and Spezzaferri [1984] in the context of predictive utility), or, within M-open frameworks, on the use of scoring rules (for example, Gneiting [2011] who considers a variety of scoring functions in the context of point prediction, or more recently Musio and Dawid [2013] who consider their use in the context of model selection). However, the cross validatory application of these rules (see, e.g. Vehtari and Lampinen [2002], Vehtari and Ojanen [2012]) is often too computationally intensive to implement, and in these circumstances it is useful to have alternative utility based metrics to employ.

In Chapter 4 we consider an example from the UK electricity generating market in which models are used to provide forecasts of extreme quantiles of price distributions. In such Value at Risk applications (see, for example, Duffie and Pan [1997] for an overview of this commonly used risk measure) it is important that the selection criteria reflects the specific application of the model. Having selected appropriate utility based scoring metrics (for further background on the quantile based losses we consider, see Cervera and Munoz [1996]) we show how the score can be incorporated directly within a score based information criterion which is then used to select the most appropriate model. As we discuss in Chapter 2, this is in contrast to many standard approaches, including the Bayes factor, which make assumptions about the specific nature of the analyst's utility which are often unrealistic in practice.

1.3.3 Modularity

The final context we consider is the situation in which models which have been built for one purpose, are 're-used' as components within larger models. In applied work, this situation often arises as a result of the importance of an audited, 'tried and tested' model being used. In some situations, this may be demanded from a 'use test' requirement in which to prevent abuse of a regulatory framework, a regulator insists that the same models should be used for internal control and external reporting (see Ong [2007]). In other circumstances, convenience and time may dictate that an analyst use what is already available in order to arrive at an acceptable solution within cost and time constraints.

As an example in energy market modelling, models for UK electricity prices are often constructed from component price models of the underlying fuels used for power generation (principally nuclear fuels, coal and gas) together with a model which forecasts demand from industrial and household consumers. A set of more deterministic relationships governing which generation options are preferred in which circumstances then allows the models to be combined into a forecasting tool for the resulting electricity price. For ease of exposition, we consider only highly stylised examples in this thesis (but see Weron [2007], Howison and Coulon [2009] and Benth and Kettler [2011] for a flavour of more detailed treatments).

In these situations we may build a number of larger 'aggregate' models from the available, smaller, component models, and wish to select the best performing. A naive approach is simply to compare the performance of the models on some future observed data. However, this may ignore richer data available to assess particular components of the model. In the electricity example above, it may be that we have a large amount of data to assess the performance of coal price models, but a more recent history only of nuclear or solar generation. In these cases, we argue in Chapter 5 that we may be able to obtain more robust and less volatile assessments of aggregate model performance by combining the assessments of component models, as an alternative approach.

We also suggest that the naive procedure of scoring the models against observed data may sometimes be sub-optimal if the researcher has a partially expressed belief in the underlying data generating mechanism, or in the degree to which future data is exchangeable

with data already observed. For example, the researcher may believe that, given market restructuring, future coal price dynamics are likely to be similar to those observed in a less regulated regime, and therefore it may be more appropriate to assess the likely future performance of the coal component of the model using data gathered from one less recent but perhaps more relevant time period, while assessing the gas component using more recent data.

While it might theoretically be possible to transform these beliefs into a more formal mixture of parametric and non-parametric assumptions (see, for example, Gutierrez-Pena and Walker [2001], Gutierrez-Pena and Walker [2005]) and deal with this situation using an M-complete formulation, in practice this is complex. We believe that modular model assessment may provide the analyst with a simpler method in which to intervene within the selection framework in order to allow additional insights to be incorporated as part of the model selection decision.

1.4 Thesis outline

In Chapter 2 we provide an overview of some of the commonly encountered Bayesian model selection approaches. We highlight important limitations in their application to problems in which specific assumptions of particular approaches are no longer valid.

The remaining chapters deal with aspects of the three contexts discussed above:

Chapter 3 discusses how we might modify Bayes factor selection when our interest is on specific marginals of a joint data generating process, rather than the full joint probability distribution. We discuss a possible approach which proceeds using a direct modification of the Bayes factor itself. We also present a related approach which seeks to reach a similar result using a more indirect adjustment of the priors on components within the models.

Chapter 4 discusses ways in which the future utility of decisions made using the chosen model can be accommodated within the model selection approach, and, in particular, the role of scoring rules in model selection. While much of the literature (Gneiting and Raftery [2007], Vehtari and Ojanen [2012]) focuses on the application of a chosen scoring rule using

cross-validatory techniques, we highlight practical difficulties often encountered when using such methods. We also explain that shortcomings with prequential predictive approaches, particularly where sample sizes are small and models need a reasonable amount of data on which to learn parameters, motivate a role for posterior predictive scores. We therefore introduce a new Bayesian score based information criterion (the *BPSIC*) which can be used to provide a bias corrected posterior score enabling the analyst to incorporate her chosen utility when comparing models.

In Chapter 5 we consider the assessment of models which act as 'components' within a larger model. Such modular assessment is potentially more flexible and allows new components to be assessed outside the larger system. We argue that it also provides one way for the analyst to intervene and incorporate her partial views on the data generating process and future exchangeability.

Chapter 6 provides a concluding summary and discussion of the themes and areas for further research.

Chapter 2

Overview of Bayesian model selection

2.1 Introductory remarks

The literature on Bayesian model selection is extensive. Kadane and Lazar [2004] and Vehtari and Ojanen [2012] provide comprehensive surveys. This chapter comments on a selection of the approaches proposed. For ease of exposition, unless otherwise stated, we consider the situation where we wish to choose between models M_1 and M_2, given a data sample, $x = \{x_1, x_2, \ldots, x_n\}$ of size n.

Each model, M_i, comprises a likelihood $f_i(y \mid \theta)$ and prior $\pi_i(\theta)$, where the dimension and support of the parameter θ may be different across models under comparison, and we denote by $p(y \mid x, M_i)$ the predictive probability density under model M_i of a future observation y given the observed data x. We refer to the true, but unknown, data generating process as $M_\star$.

We make the assumption throughout that even though our interest may be in predicting future observations, selection of a *single* model is desirable, as opposed to approaches which allow us to form a composite model from available candidates, for example, using Bayesian Model Averaging (see Geisser [1993], Draper [1995], Raftery et al. [1997])). This

might be, perhaps, for reasons of transparency and auditability of assumptions, where the analyst needs to present a single model to decision making stakeholders.

Bayesian model averaging may provide a more flexible and powerful framework for making future predictions, but the resulting formulation may be less easily interpretable than a single model would be. For example, within a single model, parameters can often be given a readily understood meaning. In a collection of models the resulting parameters can lose this meaning, with the result that the narrative structure can be lost, and it becomes more difficult for users of the model, or subject matter experts, to provide challenge to the underlying model structure and parameterisation.

2.2 A decision theoretic context

Within the context of predicting future observations, both Bernardo and Smith [1994] and the survey of established and recent approaches in Bayesian model selection in Vehtari and Ojanen [2012] frame the model selection problem in decision theoretic terms as choosing the model M_k which maximises the expected utility:

$$\bar{U}(M_k, \hat{a}_k) = \int u(M_k, \hat{a}_k, \tilde{y}) p(\tilde{y} \mid x, M_\star) d\tilde{y}, \tag{2.1}$$

where $\hat{a}_k$ is the prediction which maximises the expected utility if model M_k is used. Vehtari and Ojanen [2012] consider two cases, which they refer to as *reference predictive* and *projection predictive*. These differ according to the nature of the model for future data in which $\hat{a}_k$ is assumed to maximise expected utility.

We denote the true data generating mechanism for future observations y by $M_\star$, the utility of selecting model M_k and taking a predictive action $\hat{a}_k$ from available actions A, by $u(M_k, \hat{a}_k, \tilde{y})$. In the reference predictive approach, we have:

$$\hat{a}_k = \arg\max_{a \in A} \int u(M_k, a_k, \tilde{y}) p(\tilde{y} \mid x, M_k) d\tilde{y}, \tag{2.2}$$

that is to say, $\hat{a}_k$ is chosen to maximise expected utility were model M_k to generate future

data. By comparison, in the projection predictive approach we have

$$\hat{a}_k = \arg\max_{a \in A} \int u(M_k, a_k, \tilde{y}) p(\tilde{y} \mid x, M_\star) d\tilde{y}, \tag{2.3}$$

that is, $\hat{a}_k$ is chosen to maximise expected utility were model $M_\star$ used to generate future data.

Many common approaches can be categorised in terms of their assumptions of the extent of our knowledge (M-open, M-complete, M-closed) of the data generating mechanism, and the nature of the utility, which we now consider.

2.3 M-closed perspectives

2.3.1 The Bayes factor

In the M-closed perspective, the Bayes factor (Good [1950], Jeffreys [1961], Kass and Raftery [1995]) can be used to address the particular decision problem where our objective is to choose the true model. In this situation, any other model is equally unacceptable regardless of whether it might have been 'close enough' for a particular application. Here, our utility can be represented in a 'zero-one' form. Recalling that we denote the true model by $M_\star$, then the zero-one utility is

$$u(M_i) = \begin{cases} 1, & M_i = M_\star \\ 0, & M_i \neq M_\star. \end{cases} \tag{2.4}$$

In this situation, we now establish criteria for making the optimal decision. Let us denote by d_i, the decision to select model M_i. We then have the expected utility of d_i,

$$\bar{u}(d_i \mid x) = p(M_i \mid x). \tag{2.5}$$

To maximise expected utility, we choose d_1 if $p(M_1 \mid x) > p(M_2 \mid x)$. From Bayes' theorem,

we have

$$\frac{p(M_1 \mid x)}{p(M_2 \mid x)} = \frac{p(x \mid M_1)\pi(M_1)}{p(x \mid M_2)\pi(M_2)} = B_{12}\frac{\pi(M_1)}{\pi(M_2)}, \tag{2.6}$$

where $\pi(M_i)$ represents the prior probabilities of each model being the 'true' model, and where the factor

$$B_{12} = \frac{p(x \mid M_1)}{p(x \mid M_2)} = \frac{\int f_1(x \mid \theta)\pi_1(\theta)d\theta}{\int f_2(x \mid \theta)\pi_2(\theta)d\theta}, \tag{2.7}$$

which transforms the prior odds into the posterior odds, is called the *Bayes factor*.

It is optimal to choose M_1 if $\frac{p(x \mid M_1)\pi(M_1)}{p(x \mid M_2)\pi(M_2)} > 1$. In other words, our decision rule is choose M_1 if

$$B_{12} > \frac{\pi(M_2)}{\pi(M_1)}. \tag{2.8}$$

Jeffreys [1961] provides heuristic guidance on interpreting the Bayes factor by means of numerical values which represent differing weights of evidence in favour of one model over the other:

Table 2.1: Jeffreys [1961] scale of evidence for model selection

$\log_{10} B_{12}$	B_{12}	Evidence against H_2
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 1.5	10 to 31.6	Strong
1.5 to 2	31.6 to 100	Very Strong
> 2	> 100	Decisive

Importantly, Kadane and Dickey [1980] establish that zero-one utility is, essentially, the only utility under which the Bayes factor will be the optimal criterion – in all other cases, the expected utility depends on the data beyond the summary provided by the Bayes factor. This can be seen by considering criteria for deciding whether $\bar{u}(d_1 \mid x) > \bar{u}(d_2 \mid x)$ for arbitrary utilities u, given data x. Conditioning on the 'truth' of the model M and

assuming the true model is either M_1 and M_2, we have

$$p(M_1 \mid x)\bar{u}(d_1 \mid x, M_1) + p(M_2 \mid x)\bar{u}(d_1 \mid x, M_2) > \\ p(M_1 \mid x)\bar{u}(d_2 \mid x, M_1) + p(M_2 \mid x)\bar{u}(d_2 \mid x, M_2)$$

$$\frac{p(M_1 \mid x)}{p(M_2 \mid x)} > \frac{\bar{u}(d_2 \mid x, M_2) - \bar{u}(d_1 \mid x, M_2)}{\bar{u}(d_1 \mid x, M_1) - \bar{u}(d_2 \mid x, M_1)} = \frac{E[u(d_2) - u(d_1) \mid x, M_2]}{E[u(d_1) - u(d_2) \mid x, M_1]}. \tag{2.9}$$

For the Bayes factor alone to be a sufficient summary of the data to define the decision, we require that this fraction has no further dependence on the data. This requires each $u(d_i)$ to be constant given M_j, or in other words, up to constant factors we must have a utility of zero-one form.

These results set constraints on the range of applications to which the Bayes factor will be appropriate and beyond which Bayes factor methods may start to decrease in efficiency.

2.3.2 Bayes factors and the 'catch-up' effect

A further difficulty encountered with use of the Bayes factor is in the choice of prior distribution for the parameters within each model. The Bayes factor is demonstrably sensitive to choice of parameter prior, and determination of appropriate priors across a potentially large family of candidate models (and justification of these to auditors) is problematic. For large problems, it is usually feasible to elicit only broad characteristics of the prior. It may also be important for an analyst to present model results to, and accommodate, multiple stakeholders, each who may have a different prior. Analysis of uncertainty in the prior is therefore important to assess implications of misspecification or alternative specification.

Assuming models M_i, M_j have equal prior probability, the Bayes factor criterion that we should select model M_i in preference to M_j if $p(x \mid M_i) > p(x \mid M_j)$ can equivalently be expressed in the form

$$-\log p(x \mid M_i) < -\log p(x \mid M_j). \tag{2.10}$$

When the observed data x comprises a series of observations which we can sequence

$x_1, x_2, \ldots x_n$, then Dawid [1984] establishes a *prequential* (predictive sequential) formulation of the predictive density:

$$p(\mathbf{x} \mid M_i) = \prod_{k=1}^{n} p(x_k \mid M_i, x_1, \ldots, x_{k-1}).$$

Equivalently, the log density of the complete data can be expressed as the sum of the individual 'one step ahead' (prequential) logarithmic losses

$$\log p(\mathbf{x} \mid M_i) = \sum_{k=1}^{n} \log p(x_k \mid M_i, x_1, \ldots, x_{k-1}), \tag{2.11}$$

which gives rise to the selection criterion where we select model M_i if

$$\sum_{k=1}^{n} -\log p(x_k \mid x_1, \ldots, x_{k-1}, M_i) < \sum_{k=1}^{n} -\log p(x_k \mid x_1, \ldots, x_{k-1}, M_j). \tag{2.12}$$

In other words, selecting models based on their cumulative logarithmic loss (where we allow parameter updating to take place after each observation) is equivalent to selection based on the Bayes factor.

When we are concerned with how models will perform for future predictions, as opposed to how faithfully they have represented observed data, it is important to question the extent to which historical performance (assessed by the cumulative logarithmic loss) is representative of future, predictive, performance. In relation to the Bayes factor, van Erven et al. [2012] call this the 'catch up' effect: a model can initially perform poorly – for example, a vague prior has been used to allow initial data to have a greater influence on parameter updating – but after a period of 'training', it may start to out-perform alternative models.

The cumulative logarithmic loss may be dominated by poor performance on early observations and therefore fail to identify this change point at an early stage. We illustrate this in Section 2.3.3 below, where we show how Bayes factors which are allowed to be conditioned on a subset of training data may be preferable to the standard Bayes factor. Overcoming this shortcoming is a key motivation for our development of a posterior score criterion in Chapter 4 (Section 4.4.3).

It is important to note that the catch up effect occurs, not as a result of the zero-one utility assumption in the Bayes factor, but, rather, as a consequence of its M-closed assumption. To see this, suppose we retain the M-closed assumption, but change our decision problem to that of choosing between models M_i, $i \in I$, where our utility is linked to future predictive performance.

Bernardo and Smith [1994] provide an analysis of this situation under quadratic loss. Specifically, we suppose the utility of making a prediction, $\hat{y}_i$ of an observation using model M_i, when the actual observation is denoted by y is given by the quadratic loss:

$$u(M_i, \hat{y}_i, y) = -(\hat{y}_i - y)^2. \tag{2.13}$$

Suppose we observe data x. Under each model, M_i, the optimal prediction when using that model under quadratic loss is the mean of the predictive density, that is:

$$\hat{y}_i = E\left[y \mid x, M_i\right]. \tag{2.14}$$

We wish to select the model M_i which minimises expected loss, that is

$$\arg\min_{i \in I} \int (y - \hat{y}_i)^2 p(y \mid x) dy, \tag{2.15}$$

and under the M-closed assumption, we have

$$p(y \mid x) = \sum_{i \in I} p(M_i \mid x) p_i(y \mid x). \tag{2.16}$$

Bernardo and Smith [1994] show that this simplifies to choosing the model M_i which minimises the expression

$$(\hat{y}_i - \hat{y})^2 + \sum_{j \in I} (\hat{y}_i - \hat{y}_j)^2 p(M_j \mid x), \tag{2.17}$$

where

$$\hat{y} = \sum_{j \in I} p(M_j \mid x) \hat{y}_j, \tag{2.18}$$

and comment that in the situation where we are choosing between two models, this amounts to selecting the model with the highest posterior probability. Therefore, in this case, replacement of the zero-one utility with an alternative predictive utility also results in a choice of the Bayes factor as our decision criterion.

2.3.3 Robustness of Bayes factors to prior choice

In the previous section, we saw how the Bayes factor can be dominated by early 'training' observations, particularly where models have priors which adjust to data received. A closely related issue is that of the relative lack of robustness of the Bayes factor to the choice of prior. The Jeffreys-Lindley paradox (see Lindley [1957], Jeffreys [1961]) shows that when priors are improper and we use a Bayes factor to compare models M_1 and M_2, then it is possible to adjust the prior for model M_2 in such a way that, regardless of the data observed, model M_1 will always be preferred.

Many of the results in the robustness literature (see the comprehensive accounts in Berger et al. [1994] and Rios Insua and Ruggeri [2012]) are concerned with changes in the predictive density as a result of changes in prior. When priors are proper, these results can also be applied to the analyse the robustness of the Bayes factor to changes in prior. Berger [1990] provides an extensive review of approaches to assessing robustness, distinguishing between *local* and *global* robustness analysis. The local analysis of robustness considers the effect on posterior quantities of small perturbations in the prior, essentially studying the rate of change in posterior with the prior (see, e.g. Gustafson [1996]).

The analyst may hope that small changes in prior specification have minor impact which will further decrease as the data available for updating the model grows. This would then provide a degree of confidence in eliciting priors, or ranges of priors, which are 'good enough', and a greater degree of justification for adopting the resulting conclusions without the need for lengthy sensitivity analysis.

In the broader setting of general convergence of posterior distributions for a range of prior distributions, Gustafson and Wasserman [1995] provided an alarming finding: using the total variation metric, the supremum of the distance between posteriors resulting from

priors close in the total variation sense, almost surely *diverges* at a rate of $n^{k/2}$, where k is the dimension of the parameter space. Smith and Rigat [2012] have shown that this result was due to the coarseness of the total variation metric used, where the total variation is defined as

$$d_V(f,g) = \int_\theta \mid f(\theta) - g(\theta) \mid d\theta.$$

Using an alternative density ratio metric, the *local DeRobertis* distance, defined as

$$d^R(f,g) = \sup_{\theta,\phi} \left| \frac{f(\theta)g(\phi)}{f(\phi)g(\theta)} - 1 \right|,$$

Smith and Rigat [2012] show that, in most cases encountered in practice, where further smoothness conditions on the prior hold, the posterior distance can be bounded, and decreases in sample size.

In many cases it is also important to undertake some form of global robustness analysis, in order to provide decision makers with confidence in the the range of posterior outcomes, and therefore Bayes factors, resulting from a particular class of prior distributions.

Lavine [1991] considers a variety of classes of prior distributions useful for this purpose. Two of the most widely used classes are ϵ-*contamination* classes (Berger and Berliner [1986]) and *density ratio* classes (DeRobertis and Hartigan [1981]). For a fixed distribution π_0, and fixed value of $\epsilon \in [0,1]$, the ϵ-contamination class, Γ, relative to a family of distributions Q is defined as

$$\Gamma = (1-\epsilon)\pi_0 + \epsilon q : q \in Q. \tag{2.19}$$

Density ratio classes are defined by non-negative functions $a(\theta), b(\theta)$, where the density-ratio class S consists of the set of prior distributions with kernel densities $p(\theta)$ satisfying

$$a(\theta) \leq p(\theta) \leq b(\theta). \tag{2.20}$$

2.3.4 Assessment of Bayes factor robustness to prior choice

In this section we introduce new methods for assessing robustness to prior choice. Suppose we are interested in estimating the range of Bayes factors for comparing two models M_1 and M_2 which could result from a plausible range of prior densities, $\pi(\theta)$ around a 'base' prior, $\pi_i(\theta)$, for the parameter θ in model M_i.

A natural sub-class of the density ratio class to consider could be:

$$\Pi_i = \{\pi(\theta) : \frac{\pi(\theta)}{k} \leq \pi_i(\theta) \leq k\pi(\theta)\}, \tag{2.21}$$

for $k > 0$. For example, this could be chosen to correspond to the elicitation of likely error ranges around a subject matter expert's assessment of probability, or to reflect the range in prior opinion across multiple stakeholders: a choice of $k = 3$ would mean that the probability for any event or interval would be at most three times higher or at least a third of the quoted probability.

In Appendix A we present an algorithm (Algorithm 1) we have developed to allow us to study the extent to which the Bayes factor can vary over the range of priors within a given density ratio class, and in particular, the supremum of the Bayes factor over the density ratio class.

We now illustrate the approach with a very simple example. A normal model, $y_i \sim N(\theta, 10)$, with known standard deviation, 10, and unknown mean, θ, is estimated from a sample from the true, but unknown, data generating process with distribution $N(5, 10)$. Suppose the prior density on θ is $N(2, 5)$. After a random sample of ten observations, a posterior density of mean 5.09 and standard deviation of 2.67 is estimated. Figure 2.1 shows the prior within the density ratio family which maximises the Bayes factor where $k = 2$, (based on $N = 1,000,000$ simulations and $M = 100$ divisions) calculated using the algorithm.

Figure 2.2 shows how the ratio of the maximised Bayes factor to the Bayes factor based on the base prior changes with the sample size.

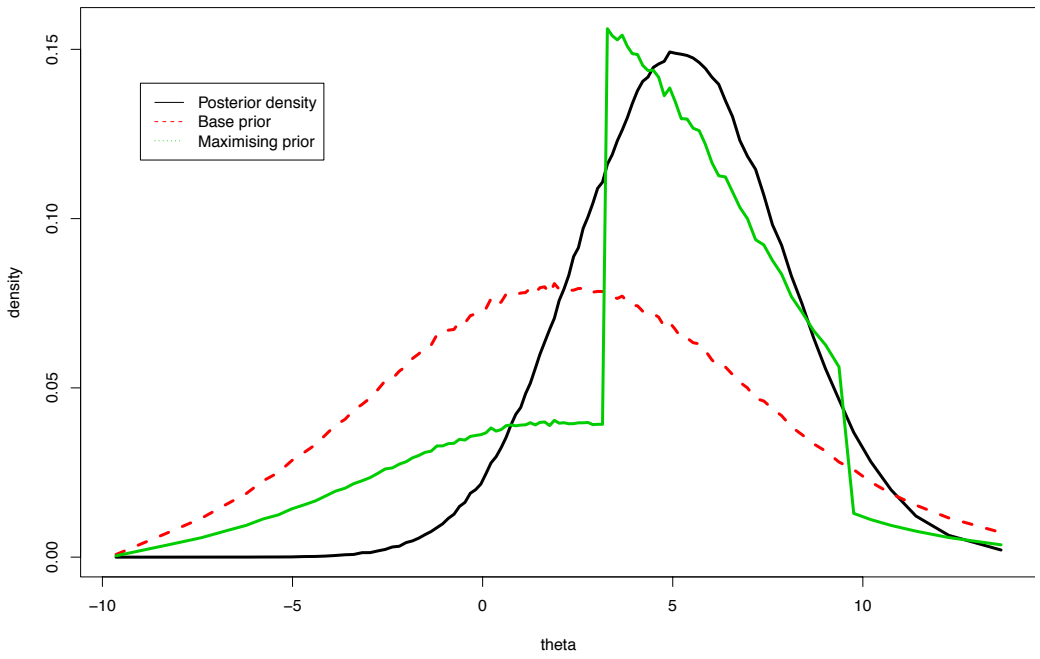

Figure 2.1: Maximising prior estimate - unknown mean, fixed sample size

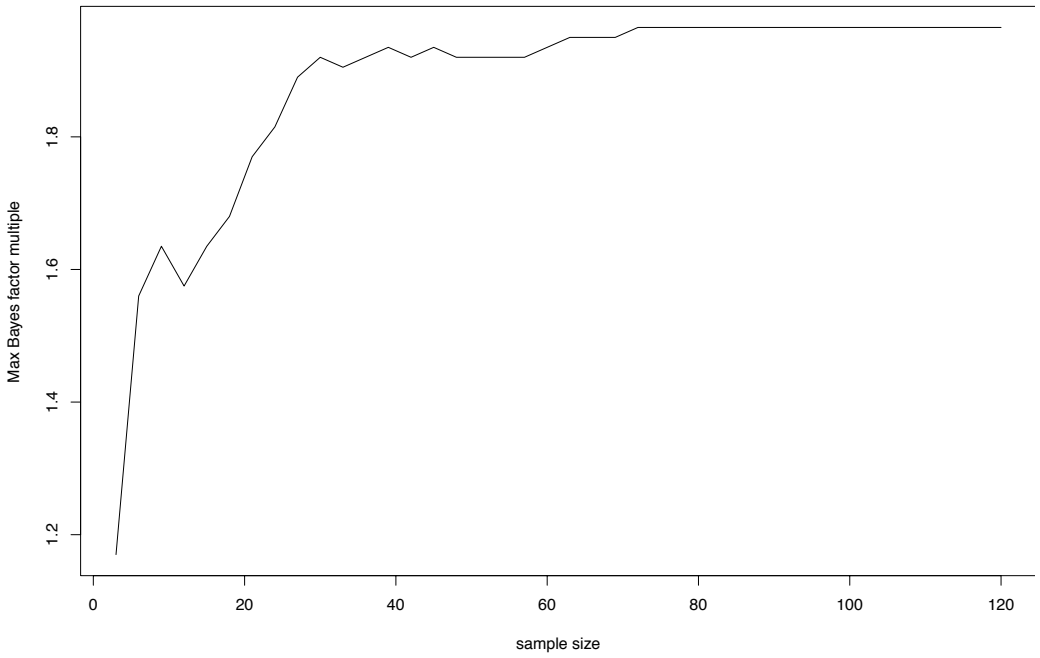

Figure 2.2: Maximum Bayes factor multiples with different sample sizes - unknown mean

In the limit, as the posterior becomes concentrated around the true mean, the full value of k (in this case 2) can be obtained as an uplift, by selection of a prior which increases its mass by a factor of k on the increasingly small interval on which the posterior concentrates its mass. The effects of the relatively poorer performance of the model in the early 'training' stages persist indefinitely, despite the model improving its predictive accuracy as parameter estimates become more accurate with increasing data.

For predictive purposes, the initial calibration of the prior is less relevant, but because the Bayes factor scores how well models did in the past, it strongly selects one model over another where these are different. Smith and Daneshkhah [2010] provide an extreme example of a model which is clearly suboptimal from the perspective of Bayes factor selection, but which performs adequately for the purpose of future prediction.

It is also interesting to compare the robustness of the Bayes factor obtained above to the robustness of the 'leave one out' Bayes factors considered at the end of Section 2.4.2. Algorithm 2 in Appendix A provides a method of doing this. For comparison, we apply this algorithm to an geometrically averaged 'leave one out' Bayes factor (which we motivate in Section 2.4.2) defined as

$$\prod_{j=1}^{k} (B_{12}(x_j \mid x_{(j)}))^{1/k}, \tag{2.22}$$

with $B_{12}(x_j \mid x_{(j)})$ representing the Bayes factor for the single observation x_j based on *updated* models incorporating data from the residual observations.

We choose an example where the true data generating process is given by $N(5, 10)$, and the model has known precision, but the prior of $N(2, 5)$ is used for the unknown mean. In this case, to compare with the previous analysis, we restrict $a(\theta), b(\theta)$ to lie within the subclass of the density ratio class given by Equation (2.20).

As can be seen from Figure 2.3, the effect of the prior being allowed to vary within the density ratio class diminishes as it is updated by the 'training sample' (that is, the sample excluding the left out value). This effect is preserved when the geometric average is taken over all possible left out values, leading to a significantly more robust measure in terms of variability within the density ratio class. The reason for this is that the performance of the

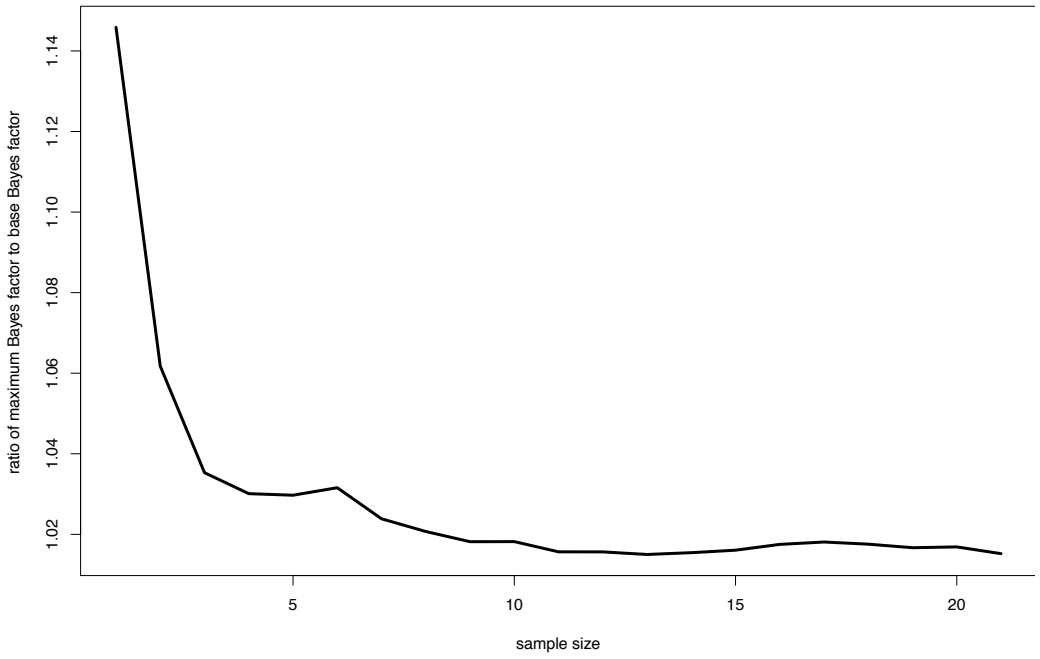

Figure 2.3: Maximum 'leave one out' Bayes factor multiples with different sample sizes - unknown mean

model is assessed on the left out values only, allowing the model to adjust its parameters on the training sample, so that the legacy of poor performance is not included within the model score.

2.3.5 Bayes factors and improper priors

An alternative approach to the problem of choosing priors is to attempt to construct 'objective' priors, which are often improper in that they do not integrate to one. These seek to represent a state of ignorance about the parameters. An analyst making use of improper priors for model parameters encounters a difficulty when using the Bayes factor, in that the values of $p(x \mid M_i)$ in Equation (2.6) are indeterminate under improper priors. In general, it will not be possible for the parameter priors under model M_i to be normalised to integrate to one. A number of modifications have been suggested to address this problem.

For example Berger and Pericchi [1996] propose a method where a *minimal training sample* such that the resulting posterior is proper is used to 'convert' the improper prior into a proper prior, and the remaining data used to compute the Bayes factor. Averaging this across the possible choices of training sample results in the *intrinsic Bayes factor*.

An alternative approach, proposed in O'Hagan [1995] and discussed further in De Santis and Spezzaferri [1999], is the *fractional Bayes factor*, defined as

$$B_{ij}^{b}(x) = \frac{q_i(x;b)}{q_j(x;b)}, \qquad (2.23)$$
$$q_i(x;b) = \frac{\int \pi_i(\theta_i) f_i(x \mid \theta_i) d\theta_i}{\int \pi_i(\theta_i) f_i(x \mid \theta_i)^b d\theta_i},$$

where b denotes a *training fraction* between zero and one, and $f_i(x \mid \theta_i), \pi_i(\theta_i)$ denote the likelihood and parameter prior for model i. Although these approaches have the benefit of 'stabilising' the improper prior, particularly when sample sizes are small and consistency properties similar to the Bayes factor can be established (see O'Hagan [1995], Ando [2010]) the choice of training fraction can be somewhat arbitrary, and particularly in the small sample case, it is unclear how much information may be lost in not using the full sample for evaluation.

2.4 M-open perspectives

In the M-open situation, the essential challenge is to find a proxy for the true model $M_\star$. Three main families of approaches have been proposed, which we consider below.

2.4.1 Posterior predictive approaches

Posterior predictive approaches (for example, Gelfand and Dey [1994], Gelfand and Ghosh [1998]) 're-use' an identical data sample $x' = x$, as a proxy for future observations under $M_\star$. The expected utility is estimated by averaging the utility of using the predictive distribution on these observations:

$$\bar{U}(M_k, \hat{a}_k) \approx \frac{1}{n} \sum u(M_k, \hat{a}_k, x_i') p(x_i' \mid D, M_\star). \qquad (2.24)$$

Posterior Bayes factors

As an extreme example of a posterior predictive approach, Aitkin [1991] proposes the *posterior Bayes factor*:

$$p(x' \mid x, M_1)/p(x' \mid x, M_2). \tag{2.25}$$

The procedure suggested is to use the complete data sample as a training sample on which to update parameters for the model. We subsequently compute the Bayes factor for the updated model against the same set of observations. Although the use of the updated model corresponds to the 'state' in which the model would be anticipated to be used for making future predictions, the contributed discussion of Aitkin [1991] identified some less desirable features of the posterior Bayes factor, most notably in using the same data twice: to estimate parameters and to assess model performance a bias is introduced which may be significant.

Goldstein [1991] constructs an extreme example of such overfitting. A random number between 1 and 1000 is selected as the parameter θ. The data value X is generated according to one of two models: $M_1 : X = \theta, M_2 : X = Z[1, 1000]$, where the distribution Z denotes a uniformly selected integer in the range 1 to 1000. In this example, the posterior Bayes factor would favour M_1 by 1000 to 1 (compared to the 'classical' Bayes factor of 1). However, the performance on the posterior Bayes factor simply reflects the extent of overfitting permitted by this model.

Lindley [1991] constructs an example where the posterior Bayes factor leads to an intransitivity in the relationship between models. In his example, four models, $M_{11}, M_{12}, M_{13}, M_{14}$ are constructed, where the posterior Bayes factor leads M_{11} to be considered more plausible than M_{12}, and M_{21} to be considered more plausible than M_{22}. However, in the composite model (defined as the model which says that the distribution is either that defined by M_{11} or M_{21}) we have that $M_{11} \cup M_{21}$ is *less* plausible than $M_{12} \cup M_{22}$. This demonstrates a lack of coherence in the approach in that it can lead to contradictory conclusions being drawn depending on the models to which it is applied.

'Held out' posterior predictive approaches

Gelfand and Dey [1994] present a framework in which model choice is based on the performance of each model's posterior predictive density (conditional on a subset of data, S_2), on another (not necessarily distinct) subset of the same data, S_1. In this sense S_2 can be regarded as a training sample, and S_1 a sample to be 'held out' for model assessment.

This permits more general predictive densities to be employed. In this case we replace the Bayes factor with a more general ratio

$$\frac{p(S_1 \mid S_2, M_1)}{p(S_1 \mid S_2, M_2)}. \tag{2.26}$$

The Bayes factor can be regarded as a special case, where we have $S_1 = x$, and $S_2 = \emptyset$. Similarly the posterior Bayes factor results from the case in which we choose $S_2 = x$ and $S_2 = x$. The choice of $S_1 = x \backslash x_i, S_2 = x_i$ results in the *pseudo Bayes factor* (Geisser and Eddy [1979]).

Such an approach not only avoids problems associated with improper priors when calculating the Bayes factor, but also has the potential to allow more general loss functions (for example replacing the posterior predictive density with a more general scoring rule (see Section 2.5) to be incorporated in the model assessment. However, it leaves open two important questions - firstly, the extent to which overlapping subsets used for model training and validation introduce bias into the assessment, and secondly the extent to which the power of the assessment is reduced by assessing performance on models conditioned on an incomplete sample of data. This approach and associated issues are closely linked to the cross-validatory approaches we now consider.

2.4.2 Cross-validatory approaches

In a non Bayesian context, Stone [1974] introduces a cross-validation procedure which assesses performance by averaging predictive performance on a subset of observed data of models built on the remaining data. The data are first partitioned into a construction

(training) sample and a validation sample. The value of an appropriate loss function comparing predictions from the model built using the construction sample to the actual held out observations in the validation sample, is averaged across all combinations of construction and validation sample.

Under the log predictive density loss function, for example, we can define the *leave one out (LOO)* cross validated score as

$$LOO = \frac{1}{n} \sum_{i=1}^{n} \log p(x_i \mid M_i, x_{(i)}), \tag{2.27}$$

where $x_{(i)}$ denotes the data x excluding the observation x_i. Alternatives using larger validation samples are clearly also possible (although see the contributed discussion in Stone [1974], suggesting that these are less likely to be optimal). Bernardo and Smith [1994], Arlot and Celisse [2010], Key et al. [1999] and Vehtari [2001] provide further examples of the application of cross validation to the estimation of general expected utilities, and more recently Vehtari and Ojanen [2012] have advocated this approach.

One practical drawback of cross-validation is that it can be computationally complex to rebuild models for multiple training samples, particularly where MCMC methods (Gelfand and Smith [1990], Gilks et al. [1996]) are used to obtain parameter estimates. This may rule it out for initial exploratory model comparison, in situations where a large number of candidate models are being evaluated.

Although computationally less expensive variants are available, for example where random sub-samples are used, there is a danger that they may omit important observations. This could be a particular danger when our aim is to assess the model on its ability to forecast tail quantiles of the distribution. Where smaller samples are used to form the training data, this may not fully reflect the model's improving performance when conditioned on the full set of available data.

2.4.3 Cross-validatory approaches and modified Bayes factors

Although it could be argued that cross-validation lacks a formal Bayesian foundation, a striking result in Bernardo and Smith [1994] shows that, at least asymptotically, related cross-validation criteria also result from Bayesian decision problems. If we are trying to predict the value of a new observation, y, from data x, we use the data itself as a proxy for the true but unknown distribution $p_T(y \mid x)$. If the observed data x is a large sample, and we define the residual data $x_{(j)} := x \setminus x_j$, then Bernardo and Smith [1994] approximate the expected utility of the action, $a(x, M_i)$, of making a prediction of y based on the observed data x, that is, they estimate:

$$\int u(y, a(x, M_i)) p_T(y \mid x) dy, \tag{2.28}$$

using the sample approximation

$$\frac{1}{n} \sum_{j=1}^{n} u(x_j, a(x_{(j)}, M_i)). \tag{2.29}$$

For example, under quadratic loss, we would look for the model, M_i which minimises

$$\frac{1}{n} \sum_{j=1}^{n} (x_j - E_{M_i}[y \mid x_{(j)}])^2, \tag{2.30}$$

where $E_{M_i}[y \mid x_{(j)}]$ denotes the mean of the predictive distribution for model M_i, conditioned on the data $x_{(j)}$. Under the loss function given by the logarithm of the predictive density, suppose decision d_i is to choose model M_i. Decision d_1 is taken if the expected utility of d_1 is greater than d_2. That is to say, if

$$\int \log p(y \mid M_1, x) p(y \mid T, x) > \int \log p(y \mid M_2, x) p(y \mid T, x) \tag{2.31}$$

then our criterion for selecting model M_1 becomes

$$\int \log \frac{p(y \mid M_1, x)}{p(y \mid M_2, x)} p(y \mid T, x) > 0. \tag{2.32}$$

Bernardo and Smith [1994] approximate this integral, arguing that as a proxy for y and x we can use a random sample from the n partitions $\left[x_{(j)}, x_j\right]$ (where $x_{(j)}$ denotes the data x with the element x_j removed) where the first term in each partition acting as a proxy for y and the second term acts as a proxy for x. By doing this, the selection criterion becomes

$$\sum_{j=1}^{k} \log \frac{p(x_j \mid M_1, x_{(j)})}{p(x_j \mid M_2, x_{(j)})} > 0,$$

and so the criterion for selecting M_1 becomes

$$\prod_{j=1}^{k} (B_{12}(x_j \mid x_{(j)}))^{1/k} > 1, \tag{2.33}$$

with $B_{12}(x_j \mid x_{(j)})$ representing the Bayes factor for the single observation x_j based on *updated* models incorporating data from the remaining observations.

They comment that one interpretation of this result is that we have a family of alternative Bayes factors: at one extreme the (classical) Bayes factor in which the model remains unchanged and is assessed on its ability to predict the full range of observations; at the other extreme the 'leave one out' Bayes factors which assess how the model predicts a single future observation when Bayesian updating is permitted, based on the remaining observations. See also Key et al. [1999] for a generalisation of these approaches to encompass alternative utility functions.

2.5 Scoring rules for model selection

Scoring rules (see, for example, the reviews in Gneiting and Raftery [2007], Gneiting [2011]) are a form of utility on a probability forecast. They provide a numerical value, or *score* $S(p, x)$ based on a forecast probability density p, and observed value x. Commonly encountered scoring rules are the *quadratic score*:

$$QS(p, x) = 2p(x) - \int p(x)^2 dx, \tag{2.34}$$

and the *logarithmic score*:

$$LogS(p, x) = \log p(x). \tag{2.35}$$

Winkler et al. [1996] and Bernardo and Smith [1994] comment on the dual role played by scoring rules in elicitation (providing an incentive for the forecaster to provide their honest, unhedged, forecast density) and also in retrospective performance assessment. More recently, Musio and Dawid [2013] advocate the direct use of scoring rules in model selection. Observations are scored based on each model's forecast, and the selection criterion becomes that of choosing the best cumulative scoring model. This is attractive as it allows the analyst to choose the appropriate utility which will be applied, in practice, to future predictions, and then to assess models using this utility.

The logarithmic and quadratic scores are examples of *strictly proper* scoring rules, whose forecast expectation is maximised if and only if the forecast density is quoted. More formally, if the forecaster believes the forecast is best represented by the probability density p, then the scoring rule S is strictly proper if for any other probability density q, we have

$$\int S(p,x)p(x)dx \geq \int S(q,x)p(x)dx, \tag{2.36}$$

with equality iff $p = q$.

Bernardo [1979] considered the additional desirable requirement of *locality*. Local scoring rules are defined to be those which only depend on the forecast density of the value of x observed, rather than the full density (that is, $S(p,x) = S(p(x),x)$). Bernardo [1979] shows that the requirement that a scoring rule be both proper and local is, under appropriate smoothness conditions on S, equivalent to requiring S to be a logarithmic scoring rule of the form $A\log p(x) + B(x)$.

One problem, in practice, with the logarithmic score is that it significantly penalises models which differ in the tails of the distribution where the log predictive density is a large negative number. If a user is more interested in performance in the body of the distribution – perhaps in terms of a model's ability to forecast typical future observations – then this may place too much emphasis on models which fit the tails accurately.

For practical model selection purposes, we argue that locality is not a crucial property (although we accept that, from a more theoretical standpoint, departure from locality may

be at odds with the likelihood principle - see Bernardo and Smith [1994] for a discussion of this point) but, rather, our central concern is in choosing a loss function which targets a particular utility. It is interesting to note that recently, other authors have discarded the need for local scoring rules. For example, Musio and Dawid [2013] advocate non-local scoring rules, albeit in their case to avoid the indeterminacy of normalising constants when using improper priors.

More generally, Dawid [2007] formulates the role of proper scoring rules and their related divergences in the context of a general decision problem where we have an outcome space X, action space A, and loss function L. If P, Q are distributions over X, Dawid [2007] defines:

- Bayes act $a_P := \arg\inf_{a \in A} L(P, a)$
- Proper scoring rule $S(x, Q) := L(x, a_Q)$
- Entropy function $H(P) := S(P, P)$
- Divergence function $d(P, Q) := S(P, Q) - H(P)$,

for $a_P, a_Q \in A$ and where we denote $L(P, a) = E^P[L(X, a)]$, $S(P, Q) = E^P[L(X, Q)]$, where $X \sim P$.

For example, if we take the logarithmic density loss function $L(x, Q) = -\log Q(x)$, the Bayes act a_P is equal to P, the associated proper scoring rule is the logarithmic scoring rule $S(x, Q) = -\log Q(x)$ and the divergence corresponds to Kullback-Leibler divergence. Alternatively, in the case of quadratic loss, the loss function $L(x, a) = (a - x)^2$, the Bayes act a_P corresponds to choosing μ_P, where μ_P is defined to be the mean of the distribution defined by P, and the associated proper scoring rule is $S(x, Q) = (\mu_Q - x)^2$.

Note that in the second example, the scoring rule is not *strictly* proper (see, e.g. Gneiting and Raftery [2007]) because the expected score will be maximised by *any* distribution sharing the same mean as the true distribution, P, neither is it local as it depends on values of the density function of Q at points other than the observed value x.

Having selected a suitable scoring rule, one method of model selection would be to compare the scores of the candidate models against the observed data. However, similar disadvantages apply to the sequential application of scoring rules to a model to those which we have noted in relation to the Bayes factor - specifically, that the cumulative score will disadvantage poorly scoring models in their early stages of parameter learning, and will focus less on the future predictive performance, and more on historical predictive performance.

One option to improve selection performance could be to adapt the techniques introduced in van Erven et al. [2012] for the case of the Bayes factor within the context of a more general cumulative scoring framework, by constructing a 'switch distribution' to describe when (in terms of sample size) it is optimal to switch from using one model to another, and from which probabilities can be obtained to select the model of highest future predictive accuracy. van Erven et al. [2012] provide examples of how it is possible to update the switch distribution based on observed data.

Cross-validatory and posterior predictive approaches might be preferred in this situation, as they are less susceptible to this effect, basing their assessments on a model which has, to a greater or lesser extent, been updated based on a subset of the data. Ideally we would use the full data on which to assess the model, although as commented previously, this introduces a bias in using the data twice. In the next section, we consider information criteria which explicitly correct for this bias.

2.6 Information criteria

We have observed that posterior predictive approaches can be desirable for two reasons. Firstly, particularly in situations where data is scarce, they assess a model in the form in which it will be used to make future predictions by allowing the analyst to update the model in full based on data received. In this way they avoid the 'catch up effect'. Secondly, they allow for flexible assessment of the performance of this model, possibly using different utility functions, by using the average performance of the model on the observed data as a proxy for the performance on the future but unknown data. However, the use of the same data twice - for parameter estimation and model assessment - introduces a bias, as

we discuss below.

Information criteria are designed to incorporate a correction for this bias. Akaike [1973] is a landmark: models are assessed on their fit in terms of Kullback-Leibler divergence (which is estimated by the AIC) to a true model, which is based on the empirical distribution of the observed data.

If we denote the true, but unknown data generating process by $f_\star$ and the pdf of two candidate models at their respective maximum likelihood estimators $\hat{\theta}$ by $f_1(x, \hat{\theta})$, $f_2(x, \hat{\theta})$, then the difference in Kullback-Leibler divergence to the true distribution:

$$KL\left(f_\star(x), f_1(x, \hat{\theta})\right) - KL(f_\star(x), f_2(x, \hat{\theta})) \tag{2.37}$$

$$= \int f_\star(x) \log f_2(x, \hat{\theta}) dx - \int f_\star(x) \log f_1(x, \hat{\theta}) dx, \tag{2.38}$$

so that by comparing the terms on the right hand side, we can assess which model is closer in Kullback Leibler divergence to the true model. Taking the expectation of one of these terms:

$$D = E_{f_\star} \int f_\star(x) \log f_i(x, \hat{\theta}) dx, \tag{2.39}$$

it can be shown (see, for example, Claeskens and Hjort [2010]) that if we estimate the true distribution $f_\star$ by the empirical distribution of the observed data x, then the estimator

$$\hat{D} = \sum_{i=1}^{n} \log f_i(x_i, \hat{\theta}) \tag{2.40}$$

will be a biased estimator of D, where the expected bias

$$E_{f_\star}[\hat{D} - D] = \mathrm{Tr}(J^{-1}K)/n + o(1/n), \tag{2.41}$$

where we denote

$$J = -E_{f_\star} \frac{\partial \log f_i(x, \theta)}{\partial \theta}, \tag{2.42}$$

$$K = Var_{f_\star} \frac{\partial^2 \log f_i(x, \theta)}{\partial \theta^2}. \tag{2.43}$$

Making the further assumption that $\mathrm{Tr}\, J^{-1}K \approx p$ where p is the number of parameters in the model (as it would be if the candidate model and true model coincided) then we can define the AIC by correcting the estimator $\hat{D}$ using the expected bias $E_{f_\star}[\hat{D} - D]$.

So we have

$$AIC = 2(\hat{D} - E_{f_\star}[\hat{D} - D]), \tag{2.44}$$

which we can express, using the above approximations, as

$$AIC = 2l_n(x, \hat{\theta}) - 2p. \tag{2.45}$$

Motivated by consideration of the Bayes factor, Schwarz [1978] introduces an alternative *Bayesian information criterion (BIC)*, which provides an asymptotic approximation to the posterior probability of a model. The resulting criterion:

$$BIC = 2l_n(x, \hat{\theta})) - p \log n \tag{2.46}$$

has a similar form to AIC, but a higher 'penalty' for the number of model parameters. For an informal derivation of the BIC, we make use of the Laplace approximation (see Tierney and Kadane [1986]) for a vector x of dimension p

$$\int e^{-Nh(x)} dx = e^{-Nh(\hat{x})} (2\pi)^{p/2} \left|\Sigma\right|^{1/2} N^{-p/2} + O(1/N), \tag{2.47}$$

where $\Sigma = \dfrac{\partial^2 h(\hat{x})}{\partial x \partial x^T}$.

We wish to approximate the integrated likelihoods which appear in the Bayes factor Equation 2.7. To approximate the likelihood $\int f_i(x \mid \theta)\pi_1(\theta)d\theta$ for model M_i, we set $N = n$ and $h(\theta) = -1/n \log f_i(x \mid \theta) + 1/n \log \pi_i(\theta)$ within the Laplace approximation equation 2.47.

Ignoring terms less than O(N), we then have

$$\int f_i(x \mid \theta)\pi_i(\theta)d\theta \approx \log f_i(x \mid \hat{\theta}) - p/2 \log N. \tag{2.48}$$

2.6.1 Comparison of AIC and BIC

Wasserman [2000] and Claeskens and Hjort [2010] discuss differences between selection using AIC and BIC. As already observed, the two criteria serve different purposes: AIC seeks to choose the model nearest in Kullback–Leibler divergence to the true data generating model; BIC seeks to select the model with highest posterior probability, and therefore provide an approximation to the model selection which would result when using the Bayes factor.

Additionally, the two criteria have different properties as assessed in terms of their *consistency* and *efficiency*.

Informally, we define (strong) consistency of a criterion as the property that, if there is exactly one model with minimum Kullback–Leibler divergence to the true data generating model, in the set of candidate models, then with probability tending to one with sample size, the criterion will select the true model. It can be shown (see Haughton [1988], Sin and White [1996] for conditions) that both AIC and BIC are consistent in these situations. However, in the situation in which there may be more than one model (for example, where models are nested) with minimum Kullback–Leibler divergence, then AIC is not consistent. Use of AIC in this situation may result in the choice of a less parsimonious model where the simpler model is, in fact, correct, and therefore it has the potential to overfit.

By contrast, we define efficiency of a criterion as the property that, where the true model lies outside the set of candidate models, then with probability tending to one with sample size, the criterion will select the model which minimises the mean squared error of prediction. It can be shown (see Sin and White [1996]) that AIC is efficient, but BIC is not efficient.

Heuristically, in light of the above, it could be argued that BIC would be a more appropriate measure to use in M-closed contexts, and AIC a more appropriate measure in M-open contexts. It is interesting that Stone [1977] shows that AIC and leave one out cross-validation are asymptotically equivalent methods. However, the possibility of overfitting using AIC in a small sample environment, where the asymptotic properties cannot

be relied upon with any confidence, may make BIC preferable, even when we do not believe any of the candidate models is the true model.

2.6.2 Extension to different utility functions

In the previous chapter we argued that Bayes factors failed to provide an accurate assessment of a model's future predictive performance. While cross-validatory techniques can be used to overcome this shortcoming, they can also be costly to implement, and information criteria, particularly where they can incorporate a wider family of utilities, can be advantageous in this regard.

It appears that most of the research on incorporating utilities within information criteria has been developed within frequentist, rather than Bayesian, settings. For example, Linhart and Zucchini [1986] provide a substantial account. In brief, given an unknown true 'operating model', F, a discrepancy function $\Delta(\theta) = \Delta(G_\theta, F)$ is chosen to assess candidate models G, based on a consideration of the aspects of importance – popular divergences used for this purpose are Kullback-Leibler divergence, Kolmogorov discrepancy and the Pearson chi-squared discrepancy – and the expected discrepancy is estimated as:

$$E^F\left[\Delta(\hat{\theta})\right] \approx E^F\left[\Delta_n(\hat{\theta}) + \mathrm{Tr}(\Omega_n^{-1}\Sigma_n)/n\right], \tag{2.49}$$

where $\hat{\theta} = \arg\min(\Delta_n(\theta))$, $\Delta_n(\theta)$ is the empirical discrepancy of the observed data, and Ω_n, Σ_n are estimators of the matrix $(\partial^2\Delta(\theta_0)/\partial\theta_i\partial\theta_j)$ and the covariance matrix $(\sqrt{n}\partial\Delta_n(\theta_0)/\partial\theta_i)$.

Claeskens and Hjort [2010] comment that considerations of the trade-off between bias and variance mean that sometimes one model is to be preferred for estimating one quantity of interest, whereas another model may be preferable for a different quantity. They also cite Hand and Vinciotti [2003] and Hansen [2005] as examples of researchers who have advocated the need to consider the 'focus' of a model. Claeskens and Hjort [2003] seek to address this through the development of the *focussed information criterion (FIC)* where the focus is on particular parameters or functions of parameters of a data generating

process.

The analysis typically proceeds (see Claeskens and Hjort [2010]) by specifying a focus parameter of interest in terms of a function of the individual model parameters (possibly distinct) in the models under consideration. The FIC is an asymptotic estimator of the mean squared error of the estimate of the parameter of interest under each model. The model with the lowest FIC is then selected.

The above techniques are frequentist in their assumptions and application, and make use of maximum likelihood estimates of parameters within each of the models under consideration. However, from a Bayesian perspective, use of point value 'plug-in' estimators and a subsequent bias correction is problematic: it fails to account for the full uncertainty expressed by the analyst's posterior distribution (see, for example, the discussion in Celeux et al. [2006]). For example, the derivation of the BIC is based on the log likelihood and is independent of the prior specification (although it could be argued that this is an advantage in cases where improper priors are to be used).

The Deviance Information Criterion (Spiegelhalter et al. [2002], Plummer [2008], Van Der Linde [2012], Spiegelhalter et al. [2014]) sought to address this by considering the posterior distribution of the data log likelihood. Denoting the posterior mean of the model parameters by $\bar{\theta}$, DIC is defined as

$$DIC = \bar{D} + p_D = D(\bar{\theta}) + 2p_D. \tag{2.50}$$

The first term, denoted by $\bar{D}$, represents the fit as defined by the posterior expectation of the deviance:

$$\bar{D}(\theta) = E_{\theta|y}\left[D(\theta)\right] = E_{\theta|y}\left[-2\log p(y \mid \theta) + 2\log f(y)\right],$$

where $f(y)$ is an arbitrary standardising term which does not affect the model comparison, and the second measures the 'complexity' of the model, defined as

$$p_D = E_{\theta|y}\left[D\right] - D(E_{\theta|y}\left[\theta\right]).$$

Vehtari [2001] suggests an extension of the DIC to cope with arbitrary utilities, u, of the form

$$\bar{U}_{DIC} = \bar{u}(E_\theta[\theta]) + 2(E_\theta[\bar{u}(\theta)] - \bar{u}(E_\theta[\theta])), \tag{2.51}$$

where $\bar{u}(\theta)$ denotes the expected utility at the parameter value θ.

However, Ando [2007] later observed that, if the function of the complexity term in DIC is to compensate for bias in the posterior estimation of the model fit, then it is incorrectly calculated. Ando [2007] introduces a Bayesian predictive information criterion (BPIC) defined as

$$BPIC = -2E_{\theta|y}[\log L(y \mid \theta)] + 2n\hat{b}_\theta, \tag{2.52}$$

with

$$n\hat{b}_\theta = E_{\theta|y}[\log(L(y \mid \theta)\pi(\theta))] - \log(L(y \mid \hat{\theta}_n)\pi(\hat{\theta}_n)) + \mathrm{Tr}(J_n^{-1}(\hat{\theta}_n)I_n(\hat{\theta}_n)) + p/2, \tag{2.53}$$

where p represents the dimension of the parameter vector θ, $\hat{\theta}_n$ the parameter value which maximises $n^{-1}\log(L(y \mid \theta)\pi(\theta))$ and where I_n, J_n are defined as follows:

$$\begin{aligned} I_n(\theta) &= \frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial(\log f_\theta(y_k) + \log\pi(\theta)/n)}{\partial\theta}\frac{\partial(\log f_\theta(y_k) + \log\pi(\theta)/n)}{\partial\theta^T}\right), \\ J_n(\theta) &= -\frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial^2(\log f_\theta(y_k) + \log\pi(\theta)/n)}{\partial\theta\partial\theta^T}\right). \end{aligned} \tag{2.54}$$

See also Zhou [2011] for further variants based on alternative estimators. However, the BPIC and these variants are limited to the logarithmic predictive density loss function, and cannot be adapted to other utilities. In Chapter 4 we consider a new information criterion which allows the incorporation of different utilities.

2.7 Chapter summary

In this chapter we have examined a number of approaches to Bayesian model selection. In the M-closed situation, where we wish to select the true model, then Bayes factor selection is optimal. However, in most practical situations where models are simply proxies for an underlying data generating process, Bayes factor selection may not represent the analyst's

utility appropriately.

In addition, the Bayes factor can fail to 'catch up' as it assesses the incoming data in a prequential fashion, so that model performance can be affected significantly by early observations at a point when the model is still updating prior estimates. This links to the overall lack of robustness of the Bayes factor to prior choice, and in this chapter we presented some new algorithms which assist in quantifying the extent of the range of Bayes factors which would result from different priors within a density ratio class.

Where our focus is on choosing a model based on the expected accuracy of its future predictions, lack of catch up is undesirable, and particularly where priors are not carefully defined, the Bayes factor lack of robustness to prior choice can make alternative approaches preferable. However, continuing to work within the M-closed framework and simply replacing the zero-one utility of 'choosing the true model' with a utility linked to prediction error of future observations does not overcome this limitation - it is a fundamental consequence of the M-closed assumption.

In light of this, it is appealing to consider the M-open approaches which have been proposed. Cross-validatory approaches can be attractive by allowing the model to be updated on a subset of data (and therefore representing the 'current state' of the model in which it will be used). At least informally, it appears that this may provide a better assessment of how the model will perform when predicting future observations. We saw how these approaches could significantly increase the robustness to initial prior choice. However, these can be computationally intensive and, when samples are small, it is not clear what information is lost by excluding observations from the training sample.

An alternative criterion, linked to minimising Kullback–Leibler divergence, results in the AIC, which in M-open contexts may be a convenient and computationally quicker way to proceed. However this is not defined in a Bayesian way, and existing generalisations (e.g. Ando [2007]) focus on the logarithmic utility only.

If we wish to select models on more general utilities, then it is possible to define an appropriate scoring rule linked to this utility. The selection approach then becomes one of

scoring the models based on data received, with the analyst selecting the highest scoring model. In this way, if we assume exchangeability of future observed data with the historical observations, the model which is selected can be judged to provide the analyst with the greatest expected future utility.

However, when carried out in a prequential fashion, which is the natural way to proceed, the 'catch up' phenomenon will also reduce the effectiveness of an approach based on comparing cumulative scores, and this may result in the analyst selecting a model with an initially better 'track record'. While it would be possible to discard scores on initial observations, the decision on the size of the hold out sample would seem arbitrary.

In Chapter 4 we will argue that, ideally, when we are interested in more general utilities, it is useful to be able to tailor information criteria to specific scores of interest. In this way, we are able to use the full data on which to refine parameter estimates while, using an appropriate bias correction term, also assessing its performance on the data observed. We present a new Bayesian Posterior Score Information Criterion for this purpose.

In some cases, however, the full machinery of an M-open approach may not be necessary to draw appropriate conclusions, and departures from the M-closed assumption may not be drastic enough – at least on some aspects of the modelled variables – to discard analysis using Bayes factors. In the next chapter, we consider simpler modifications to the Bayes factor which may be used, particularly when our interest is on a subset of the variables within a model.

Chapter 3

Modified Bayes factors for the selection of marginal distributions

3.1 Motivating Examples

In the previous chapter, we observed that the derivation of the Bayes factor made use of the assumption that one of the models under consideration was the true data generating process for the full joint distribution. However, we also noted that consistency properties of Bayes factor selection meant that, even when the true model was not a member of the family of models entertained, it would asymptotically select the model closest in Kullback–Leibler divergence to that model.

In many applications we may often be interested only in a subset of relationships, or in particular aspects of the distribution originally modelled. Perhaps, when viewed in terms of the marginal distribution on these variables alone, the assumption that one of the models under consideration is close to the true distribution may not be unreasonable. However, the analyst may be aware that model performance in terms of its predictions of other, 'nuisance' variables (perhaps introduced for modelling convenience) may depart in a more obvious way from the true data generating process.

In this situation, there is a danger that the Kullback–Leibler divergence is dominated by

divergence on variables of less interest, and therefore using the unmodified Bayes factor may be less effective. In this chapter we therefore investigate modifications to standard Bayes factor selection which can be applied where our interests in the model are qualified in these ways.

In Section 1.3.1, we considered the situation when an analyst was interested in the dependence of price, V, on UK gas demand, D, but had chosen to build a large Bayesian network to represent this, because it allowed a series of intermediate relationships to be more readily modelled, understood and elicited.

Suppose that, to do this, the network structure introduces a number of intermediate dependencies through modelled variables $X = (X_1, X_2, \dots X_n)$, where we denote the parents of X_i by $Pa(X_i)$, and where the demand D is represented by X_1, price V is represented by X_n, and n is large. We denote the true data density by p, and the modelled density by q. Assuming the analyst's interest is in the full joint distribution, and her utility is given by the log predictive density, she may wish to choose the model with the greatest expected log score

$$\int \log q(x)p(x)dx, \tag{3.1}$$

or equivalently the model with the smallest Kullback Leibler divergence to the true data density:

$$KL(p, q) := \int \log \frac{p(x)}{q(x)} p(x)dx. \tag{3.2}$$

If, as is the case here, her interest, instead, lies either (a) in forecasting the marginal distribution of the variable X_n, or (b) in forecasting the conditional distribution $X_n \mid X_1$, then, supposing that her utility is given by the log predictive utility only for these elements, this will correspond to a *different* selection criterion, namely choosing the model which minimises (a) the *marginal* Kullback-Leibler divergence:

$$K_{X_n}(p, q) := KL(p(x_n), q(x_n)). \tag{3.3}$$

or (b) the *conditional* Kullback-Leibler divergence

$$K_{X_n|X_1}(p,q) := E^p[K_{X_n}(p(x_n \mid x_1), q(x_n \mid x_1))]. \tag{3.4}$$

However, Bayes factor comparison will automatically include the model's performance on all other relationships, regardless of whether they are pertinent to the decision maker's utility. For example, in a simple case where the variables are independent, we have

$$KL(p,q) = \sum_{i=1}^{n} K_{X_i}(p,q), \tag{3.5}$$

or more generally we will have:

$$KL(p,q) = \sum_{i=1}^{n} K_{X_i|Pa(X_i)}(p,q), \tag{3.6}$$

so that the good performance of a model on the marginal or conditional distributions of interest will be disguised by poor performance on the other marginals or conditionals which are introduced by the model. If, for example, variable X_i for some $1 < i < n$ has little or no influence over the marginal of interest, but has been poorly modelled, then its introduction will have little impact on the performance of the model on the marginal of interest, but it will have a larger negative impact on the overall model score.

The exact impact of this will depend on the extent to which data contains 'unusual' observations, but for high dimensional settings we can reasonably expect outliers on *some* dimension. In big data contexts, therefore, outlying observations on aspects of the model in which we are not interested can have a distorting impact on decisions taken if we use the logarithmic score across the full joint distribution.

Example 1

To illustrate this, first we consider the following basic example. We compare two models M_1, M_2 for the joint distribution of two variables X and Y, where the probability density function for model M_i is given by $p_i(x,y)$. For each model we assume independence of the variables, with a representation $p_i(x,y) = f_i(x)g_i(y)$. In Section

2.3.2, we saw that we can regard Bayes factor selection as equivalent to evaluating the prequential logarithmic scores.

If we choose to score the models based on the performance on the joint log density, we have $\log p_i(x, y) = \log f_i(x) + \log g_i(y)$. If our focus of interest is on variable Y alone, we *should* compare only the log densities on y, that is $\log g_i(y)$. For example, using the joint density will result in incorrect selection of model M_1 in the situation that $\log g_2(y) > \log g_1(y)$, but $\log f_1(x) > \log f_2(x) + \log g_2(y) - \log g_1(y)$.

Example 2

We now consider a more realistic situation in which two models for variables X and Y are compared. This time we model X and $Y \mid X$. We show how we would expect to prefer one model as a model for X and $Y \mid X$, but expect to prefer the other model as a model for Y.

Lemma 1 *Suppose the true data generating process is given by $M_\star : Y \sim N(m, s^2)$. Then the expected log score of model $M_1 : Y \sim N(\mu, \sigma^2)$ is*

$$E\left[S_1\right] = -\frac{1}{2}\left[\log 2\pi\sigma^2 + \frac{1}{\sigma^2}(s^2 + (m-\mu)^2)\right]. \tag{3.7}$$

For fixed μ, $E[S_1]$ is maximised when $\sigma^2 = s^2 + (m-\mu)^2$.

Proof.

$$\begin{aligned} E\left[S_1\right] &= -\frac{1}{2}\left[\log 2\pi\sigma^2 + E\left[\frac{(x-\mu)^2}{\sigma^2}\right]\right] \\ &= -\frac{1}{2}\left[\log 2\pi\sigma^2 + \frac{1}{\sigma^2}E\left[((x-m) + (m-\mu))^2)\right]\right] \\ &= -\frac{1}{2}\left[\log 2\pi\sigma^2 + \frac{1}{\sigma^2}(s^2 + (m-\mu)^2)\right]. \end{aligned} \tag{3.8}$$

The maximum value is obtained by differentiating this expression with respect to σ. ■

Suppose the true data generating process is bivariate normal with correlation r. In this case, for the true data generating process, $M_\star$, we have:

$$\begin{aligned} X &\sim N(m_X, s_X^2) \\ Y &\sim N(m_Y, s_Y^2) \\ Y \mid X &\sim N(m_Y + (s_Y/s_X)r(X - m_X), (1-r^2)s_Y^2). \end{aligned} \tag{3.9}$$

Further suppose we choose to model this using model M_1 for which we assume the following:

$$\begin{aligned} X &\sim N(\mu_X, \sigma_X^2) \\ Y &\sim N(\mu_Y, \sigma_Y^2) \\ Y \mid X &\sim N(\mu_Y + (\sigma_Y/\sigma_X)\rho(X - \mu_X), (1-\rho^2)\sigma_Y^2). \end{aligned} \tag{3.10}$$

Theorem 2 *The expected scores on X, Y, $Y \mid X$ for model M_1 are as follows:*

$$\begin{aligned} E\left[S_1^X\right] &= -\frac{1}{2}\left[\log 2\pi A^X + \frac{1}{A^X}(B^X + C^X)\right] \\ A^X &= \sigma_X^2, B^X = s_X^2, C^X = (m_X - \mu_X)^2 \end{aligned} \tag{3.11}$$

$$\begin{aligned} E\left[S_1^Y\right] &= -\frac{1}{2}\left[\log 2\pi A^Y + \frac{1}{A^Y}(B^Y + C^Y)\right] \\ A^Y &= \sigma_Y^2, B^Y = s_Y^2, C^Y = (m_Y - \mu_Y)^2 \end{aligned} \tag{3.12}$$

$$\begin{aligned} E\left[S_1^{Y|X}\right] &= -\frac{1}{2}\left[\log 2\pi A^{Y|X} + \frac{1}{A^{Y|X}}(B^{Y|X} + C^{Y|X})\right] \\ A^{Y|X} &= (1-\rho^2)\sigma_Y^2, B^{Y|X} = (1-r^2)s_Y^2, \\ C^{Y|X} &= (\frac{\rho\sigma_Y}{\sigma_X}(\mu_X - m_X) + (m_Y - \mu_Y))^2 + (rs_Y - \frac{\rho s_X \sigma_Y}{\sigma_X})^2. \end{aligned} \tag{3.13}$$

Proof. The results for the scores on X and Y follow directly from Lemma 1. For the expected score on $Y \mid X$, we have

$$\begin{aligned} E_S^{Y|X} &= E[E_S^Y \mid X = x] \\ &= -\frac{1}{2}E[\log 2\pi(1-\rho^2)\sigma_Y^2 + \frac{1}{(1-\rho^2)\sigma_Y^2}((1-r^2)s_Y^2 + C)] \\ C &= E[m_Y + \frac{s_Y}{s_X}r(X - m_X) - \mu_Y - \frac{\sigma_Y}{\sigma_X}\rho(X - \mu_X)]^2. \end{aligned} \tag{3.14}$$

Observing that $E[aX + b]^2 = (am_X + b)^2 + a^2 s_X^2$, we have:

$$\begin{aligned} C &= E[(\frac{s_Y}{s_X}r - \frac{\sigma_Y}{\sigma_X}\rho)X + (m_Y - \mu_Y + \frac{\sigma_Y}{\sigma_X}\rho\mu_X - \frac{s_Y}{s_X}rm_X)]^2 \\ &= ((\frac{s_Y}{s_X}r - \rho\frac{\sigma_Y}{\sigma_X})m_X + (m_Y - \mu_Y + \frac{\sigma_Y}{\sigma_X}\rho\mu_X - \frac{s_Y}{s_X}rm_X))^2 + (\frac{s_Y}{s_X}r - \rho\frac{\sigma_Y}{\sigma_X})^2 s_X^2 \\ &= (\frac{\rho\sigma_Y}{\sigma_X}(\mu_X - m_X) + (m_Y - \mu_Y))^2 + (r\frac{s_Y}{s_X} - \frac{\rho\sigma_Y}{\sigma_X})^2 s_X^2 \end{aligned} \tag{3.15}$$

from which the result follows. ■

This has an important consequence. Suppose we are comparing the expected scores of two models M_0 and M_1 following the specification in Equation 3.10. Then if we set the values of σ_X and σ_Y to be equal in both models, but in model M_1 we set $\mu_Y = m_Y - \frac{\rho\sigma_Y}{\sigma_X}(\mu_X - m_X)$, this will maximise the expected score on $Y \mid X$ (and therefore also on the joint distribution of X and Y), whereas if in model M_1 we set $\mu_Y = m_Y$, this will maximise its expected score on variable Y.

This means that using Bayes factor selection we will prefer the two different models in different situations. If we build a Bayesian network as in Figure 3.1, where our interest is in variable Y, scoring based on the represented nodes X and $Y \mid X$ (in other words, scoring the full network) is not appropriate if we aim to select the best model for Y.

3.2 Restricted Bayes factors

In order to avoid the problems with the use of the full Bayes factor illustrated in the previous section, we next consider a *restricted* Bayes factor which is designed to focus on only those relationships which are of interest to us.

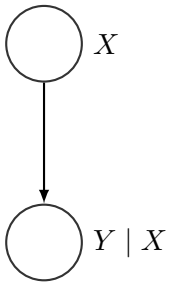

Figure 3.1: Two node Bayesian network

We consider multivariate data $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$, and suppose our aim is to select models when our interest is in a subset of *independent* relationships (marginal distributions) $\mathbf{R}^\star = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_p\}$ of a complete set of independent relationships $\mathbf{R} = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_q\}$ $(p < q)$, complete in the sense that the joint probability density of the full data $\mathbf{X}$ is equal to the product of all densities in $\mathbf{R}$. For example, as in the previous section, we might be interested only in certain nodes or conditional dependence relationships in a Bayesian network.

We denote by $p(R_i^n)$ the probability assigned by a model to the relationship $\mathbf{R}_i$ with values of the variables occurring in $\mathbf{R}_i$ taken from the observation $\mathbf{x}_n$. For example, where we observe bivariate data and $\mathbf{R}^\star$ contains the single conditional relationship $\mathbf{R}_1 = X \mid Y$, then $p(R_1^2)$ is the conditional probability $p(X = x_2 \mid Y = y_2)$.

Define the $M - \mathbf{R}^\star$ closed assumption as that in which one of the models is the 'correct' model for the distributions contained in $\mathbf{R}^\star$ (where $\mathbf{R}^\star = \mathbf{R}$ this corresponds to the M-closed setting of Bernardo and Smith [1994]). Suppose the utility of decision d_i to choose model M_i when the correct model for $\mathbf{R}^\star$ is M_j, is given by $U(d_i, M_j)$.

To select M_1, we should have:

$$\sum_{j=1}^{2} p(M_j \; correct \; for \; \mathbf{R}^\star \mid \mathbf{x}) E[U(d_1, M_j) \mid \mathbf{x}] > \sum_{j=1}^{2} p(M_j \; correct \; for \; \mathbf{R}^\star \mid \mathbf{x}) E[U(d_2, M_j) \mid \mathbf{x}],$$

or equivalently

$$\frac{p(M_1 \; correct \; for \; \mathbf{R}^\star \mid \mathbf{x})}{p(M_2 \; correct \; for \; \mathbf{R}^\star \mid \mathbf{x})} > \frac{E[U(d_2, M_2) \mid \mathbf{x}] - E[U(d_1, M_2) \mid \mathbf{x}]}{E[U(d_1, M_1) \mid \mathbf{x}] - E[U(d_2, M_1) \mid \mathbf{x}]}. \tag{3.16}$$

By extension of the argument in Kadane and Dickey [1980] we can see that in order for the right hand side to be independent of the observed data, we require the ratio on the right hand side not to depend on $\mathbf{x}$. This in general will only be possible for constant utilities which depend only on choosing the 'right' model. For now, let us assume that we do indeed have such a utility. If we assume that the two models have the same prior probability, we have

$$\frac{p(M_1 \; correct \; for \; \mathbf{R}^\star \mid \mathbf{x})}{p(M_2 \; correct \; for \; \mathbf{R}^\star \mid \mathbf{x})} = \frac{p(\mathbf{x} \mid M_1 \; correct \; for \; \mathbf{R}^\star)}{p(\mathbf{x} \mid M_2 \; correct \; for \; \mathbf{R}^\star)}. \tag{3.17}$$

We call the right hand term of this expression the ***R***-restricted Bayes factor*.

For the standard log Bayes factor, we can calculate in a sequential fashion, employing the prequential (Dawid [1984]) representation

$$\log \frac{p(\mathbf{x} \mid M_1)}{p(\mathbf{x} \mid M_2)} = \sum_{i=1}^{n} (\log p(\mathbf{x}_i \mid M_1, \mathbf{x}_1, \ldots \mathbf{x}_{i-1}) - \log p(\mathbf{x}_i \mid M_2, \mathbf{x}_1, \ldots \mathbf{x}_{i-1})). \tag{3.18}$$

Given the completeness and independence assumptions on $\mathbf{R}$ we also have that the log Bayes factor can be expressed as

$$\log \frac{p(\mathbf{x} \mid M_1)}{p(\mathbf{x} \mid M_2)} = \sum_{i=1}^{n} \sum_{j=1}^{p} (\log p(R_j^i \mid M_1, \mathbf{x}_1, \ldots \mathbf{x}_{i-1}) - \log p(R_j^i \mid M_2, \mathbf{x}_1, \ldots \mathbf{x}_{i-1})). \tag{3.19}$$

For the $\mathbf{R}^\star$-restricted Bayes factor, using the assumption that the relationships in $\mathbf{R}^\star$ are independent we have

$$\log \frac{p(\mathbf{x} \mid M_1 \; correct \; for \; \mathbf{R}^\star)}{p(\mathbf{x} \mid M_2 \; correct \; for \; \mathbf{R}^\star)} = \tag{3.20}$$
$$\sum_{i=1}^{n} \sum_{j=1}^{p} (\log p(R_j^i \mid M_1, \mathbf{x}_1, \ldots \mathbf{x}_{i-1}) - \log p(R_j^i \mid M_2, \mathbf{x}_1, \ldots \mathbf{x}_{i-1})).$$

Note that the conditioning is on all previous observations on all variables, not just those contained within relationships in $\mathbf{R}^\star$.

3.3 Using vague priors on parameters of low interest to robustify model selection

We saw in Section 3.21 that where only a subset of variables are of interest to the analyst, it may be more appropriate to compute a restricted, rather than full, Bayes factor. In some cases, it may not be possible to do this directly. This might be the case, for example, when MCMC techniques (Gelfand and Smith [1990], Gilks et al. [1996]) or 'off the shelf' software routines are being used (for example, Lunn et al. [2000], Madsen et al. [2005]) and it may be possible to compute the full predictive density only.

Comparing the expressions in Equations (3.20) and (3.19) we can see that the difference between using the log Bayes factor and the log $\mathbf{R}^\star$-restricted Bayes factor is equal to

$$\sum_{i=1}^{n} \sum_{j=p+1}^{q} (\log p(R_j^i \mid M_1, \mathbf{x}_1, \ldots \mathbf{x}_{i-1}) - \log p(R_j^i \mid M_2, \mathbf{x}_1, \ldots \mathbf{x}_{i-1})). \tag{3.21}$$

This term is the difference between the log predictive distributions of the two models over the nuisance variables. Where we can arrange for this term to be relatively small, then using the standard Bayes factor will provide a similar answer to the restricted Bayes factor.

One situation in which this can be achieved is in a hypothesis testing scenario in which we are comparing two models with the same likelihood but whose parameters are expressed in terms of priors which are zero on regions of the parameter space (and are typically disjoint). We would expect the difference between cumulative log scores on these variables to diverge; on the other hand if we loosen the priors of the two models on the nuisance variables to allow them to take the unrestricted range of values and those priors are allowed to adapt to incoming data, we might expect that the difference in Equation 3.21 tends to a limit for these variables.

By doing this, the Bayes factor becomes dominated by the sharper priors retained on those variables of interest, so the sum of the difference in log scores on variables will tend to a finite limit, whereas on the variables of interest, the difference in log scores on these

will tend to infinity. The net effect is that the model selection, at least asymptotically, favours the model which is 'correct' on the variables of interest.

To formalise this approach, we prove that under a 'loose' prior (loose in the sense that its support contains the value of the parameter which minimises the Kullback Leibler divergence to the true model), the difference in cumulative log scores between two models sharing the same likelihood, but having different priors, tends to a constant limit.

Lemma 3 *Suppose we have two models M_1 and M_2 which share the same likelihood for a series of n future observations x^n, where we denote the likelihood by $p(x^n \mid \theta)$. Suppose the two models have different priors for the parameters, which we denote by $\pi_i(\theta)$, and which include within their support the parameter value $\hat{\theta}_0$ which minimises the Kullback Leibler divergence of $p(x \mid \theta)$ to the true density. Then the difference in cumulative log scores can be approximated as*

$$\log \frac{\int \pi_1(\theta) p(x^n \mid \theta) d\theta}{\int \pi_2(\theta) p(x^n \mid \theta) d\theta} = \log \frac{\pi_1(\hat{\theta}_0)}{\pi_2(\hat{\theta}_0)} + O(1/N). \tag{3.22}$$

Proof. We use the Laplace approximation (Tierney and Kadane [1986]) for a vector x of dimension p

$$\int e^{-Nh(x)} dx \approx e^{-Nh(\hat{x})} (2\pi)^{p/2} |\Sigma|^{1/2} N^{-p/2} + O(1/N), \tag{3.23}$$

where $\Sigma = \dfrac{\partial^2 h(\hat{x})}{\partial x \partial x^T}$.

We wish to approximate the integrated likelihoods which appear in the Bayes factor equation 2.7. To approximate the likelihoods $\int f_i(x \mid \theta)\pi_1(\theta)d\theta$ for model M_i, we set $N = n$ and $h_i(\theta) = -1/n \log f_i(x \mid \theta) + 1/n \log \pi_i(\theta)$ within the Laplace approximation equation 3.23. Cancelling terms which are identical, we have that

$$\begin{aligned} \log \frac{\int \pi_1(\theta) p(x^n \mid \theta) d\theta}{\int \pi_2(\theta) p(x^n \mid \theta) d\theta} &= \log \int \pi_1(\theta) p(x^n \mid \theta) d\theta - \log \int \pi_2(\theta) p(x^n \mid \theta) d\theta \\ &= -n h_1(\hat{\theta}) + n h_2(\hat{\theta}) + O(1/N) \\ &= log \frac{\pi_1(\hat{\theta}_0)}{\pi_2(\hat{\theta}_0)} + O(1/N) \end{aligned} \tag{3.24}$$

■

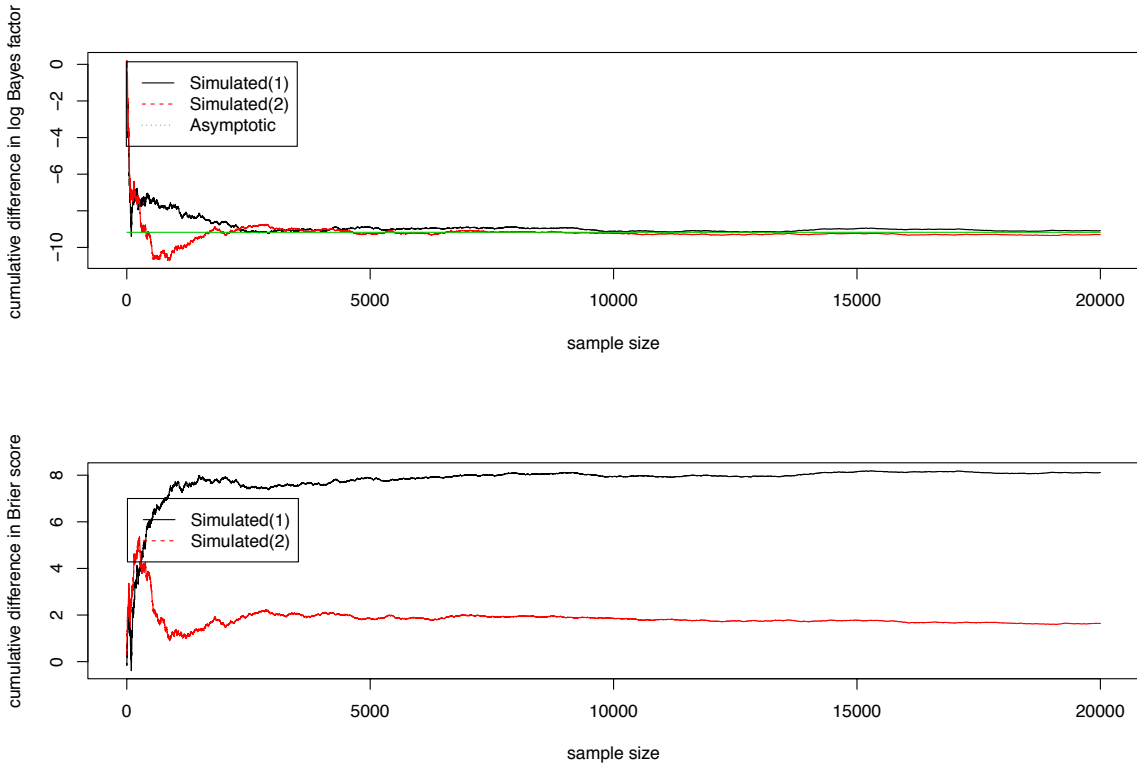

Figure 3.2: Large sample behaviour of cumulative log Bayes factor (cumulative log score) and cumulative Brier score for beta/binomial models M_0, M_1. In this case two sets of data were generated from a 'true' binomial model with parameter $p = 0.5$. Priors for M_1, M_2 were $Be(\alpha, \beta)$ with $M_0 : \alpha = 21, \beta = 7$, $M_1 : \alpha = 45, \beta = 9$. Note that, unlike the Brier score, the difference in cumulative log scores converges to a limit independent of the realisation of the random sample.

This property is worthy of comment. It implies that the difference between log scores of the two models tends to a constant limit, irrespective of the order in which data are received, or whether one of the models is 'true'. In the prequential interpretion (see Dawid [1984]) of the log score as the sum of one step ahead scores, it means that a particularly unlikely observation under one model is compensated for exactly by a relative improvement in that model's future forecasting performance through parameter updating, as measured by the the future log score difference.

It should be contrasted with the behaviour of other proper scoring rules, for example the Brier score for which the limiting behaviour *is* dependent on the specific realisation of observed data, as illustrated in Figure 3.2. As examples, we have the following results for the beta/Bernoulli and normal/inverse Wishart distributions (which are also derived

from first principles in Appendix B):

Example 4 *Suppose the true data generating process of univariate binary valued data has mean μ, and that we compare two conjugate Beta/Bernoulli models, where the prior on model M_i is parameterised by α_i, β_i. Assuming a data sample of size n from the true data generating process is observed, then as $n \to \infty$, the difference in log scores tends to a constant limit of*

$$\log \frac{B(\alpha_0, \beta_0)}{B(\alpha_1, \beta_1)} + (\alpha_1 - \alpha_0) \log \mu + (\beta_1 - \beta_0) \log(1 - \mu),$$

where $B(\alpha, \beta)$ denotes the beta function $\dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$.

In particular, if the two models share the same prior mean, $\hat{\mu}$, but different prior variance, assuming large enough α_i, β_i, say $\alpha_i = K_i\alpha, \beta_i = K_i\beta$ with K_i large, but $K_i \ll N$,

$$\log \frac{p_1(x)}{p_0(x)} \to (K_0 - K_1) \log \left(\left(\frac{\hat{\mu}}{\mu} \right)^{\alpha} \left(\frac{1 - \hat{\mu}}{1 - \mu} \right)^{\beta} \right) - \frac{1}{2} \log \left(\frac{K_0}{K_1} \right). \tag{3.25}$$

Example 5 *Suppose the true data generating process of multivariate data of dimension d has mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and that we compare two conjugate normal inverse Wishart models, where the prior on model M_i is parameterised by $\boldsymbol{\mu}_i, \kappa_i, \boldsymbol{\Lambda}_i, \nu_i$. Assuming a data sample of size n from the true data generating process is observed, then as $n \to \infty$, the difference in log scores tends to a constant limit of*

$$\frac{1}{2}(d \log \frac{\kappa_0}{\kappa_1} + \nu_0 \log |\boldsymbol{\Lambda}_0| - \nu_1 \log |\boldsymbol{\Lambda}_1| + (\nu_1 - \nu_0)(\log |\boldsymbol{\Sigma}| + d \log 2))$$

$$+\frac{1}{2} \operatorname{Tr}((\boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_0 + \boldsymbol{D}_1 - \boldsymbol{D}_0)\boldsymbol{\Sigma}^{-1}) + \sum_{i=1}^{d} (\log \Gamma(\frac{\nu_1 + 1 - i}{2}) - \log \Gamma(\frac{\nu_0 + 1 - i}{2})),$$

where $\boldsymbol{D}_i$ is defined as $\kappa_i(\boldsymbol{\mu} - \boldsymbol{\mu}_i)(\boldsymbol{\mu} - \boldsymbol{\mu}_i)^T$.

We illustrate this in Figures 3.3 and 3.4 in the case of two simulated data sets from a bivariate model, where the true data generating process is bivariate normal (that is, 'inside' the family), and bivariate t-distributed (that is, 'outside' the family). All

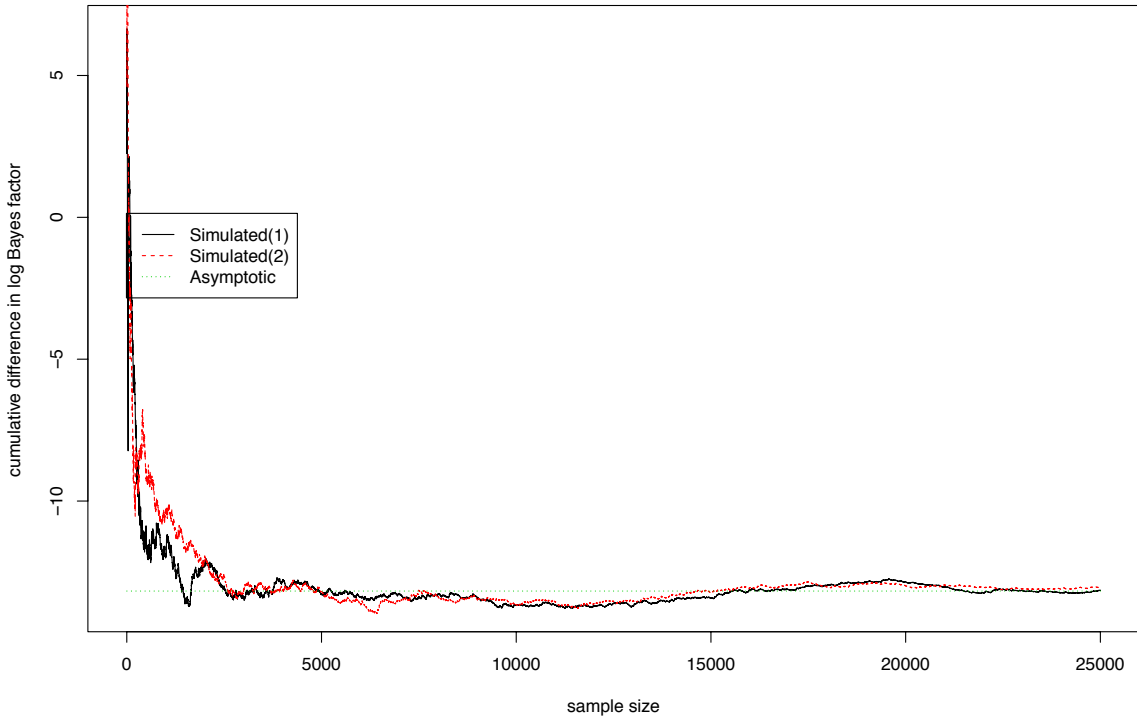

Figure 3.3: Large sample behaviour of difference in cumulative log Bayes factor (cumulative log score) for multivariate normal models, M_0, M_1. In this case the models contain the true data generating process as a specific case, and two sets of data are generated from a 'true' bivariate normal generating process with mean $\binom{3}{2}$ *and covariance matrix* $\left(\begin{smallmatrix}5 & 3\\ 3 & 5\end{smallmatrix}\right)$. *Priors for* M_1, M_2 *were* $NIW(\mu, \kappa, \mathbf{\Lambda}, \nu)$ *with* $M_0 : \mu = \binom{1}{5}, \kappa = 10, \mathbf{\Lambda} = \left(\begin{smallmatrix}6 & 3\\ 3 & 6\end{smallmatrix}\right), \nu = 3$, $M_1 : \mu = \binom{5}{3}, \kappa = 2, \mathbf{\Lambda} = \left(\begin{smallmatrix}2 & 1\\ 1 & 2\end{smallmatrix}\right), \nu = 9$.

simulations were generated using the R packages *mvtnorm* (Genz and Bretz [2009]) and *tmvtnorm* (Wilhelm and Manjunath [2012]).

This result allows us to decompose the Bayes factor into a Bayes factor on the nuisance variables (which tends to a constant limit) and a Bayes factor of the variables of interest conditional on the nuisance variables (which we would expect to increase with increasing sample size in favour of the model with least Kullback Leiber divergence to the true model on the variables of interest). Specifically, suppose the variables of interest are denoted by $\mathbf{y} := \mathbf{x}^1, \mathbf{x}^2, \ldots \mathbf{x}^m$, with the nuisance variables denoted by $\mathbf{z} := \mathbf{x}^{m+1}, \mathbf{x}^{m+2}, \ldots \mathbf{x}^n$.

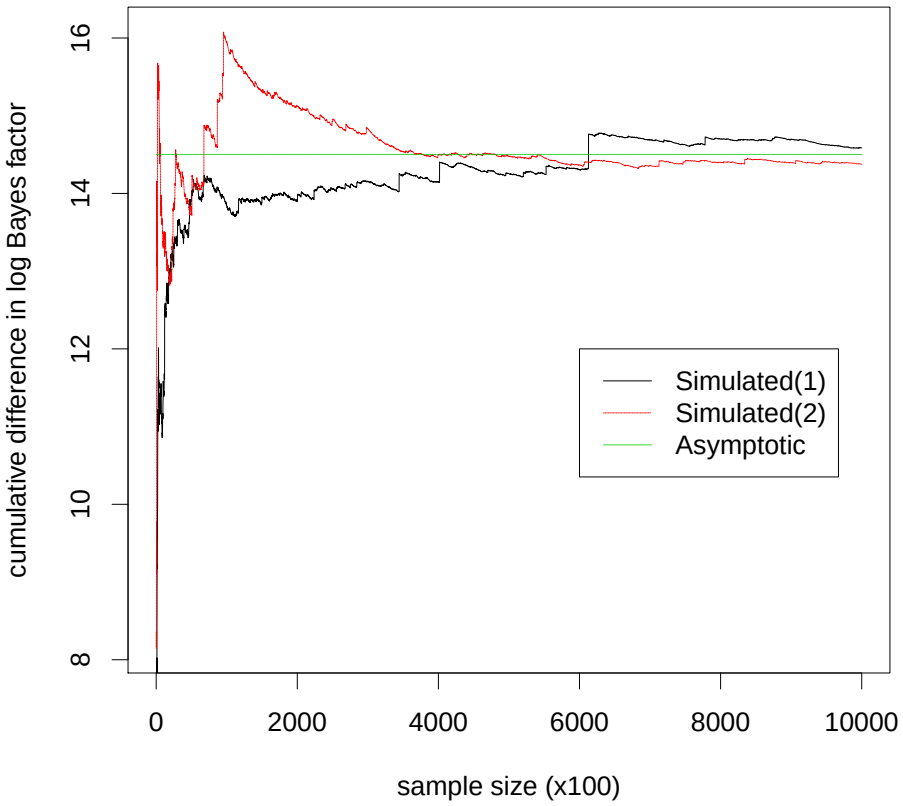

Figure 3.4: Large sample behaviour of difference in cumulative log Bayes factor (cumulative log score) for multivariate normal models, M_0, M_1. In this case the true data generating process lies outside the modelled distributions, and to simulate this two sets of data are generated from a 'true' bivariate t-distribution generating process with mean $\binom{3}{2}$, $df = 3$ and covariance matrix $df/(df-2)\left(\begin{smallmatrix}5 & 3\\ 3 & 5\end{smallmatrix}\right)$. Priors for M_1, M_2 were $NIW(\mu, \kappa, \mathbf{\Lambda}, \nu)$ with $M_0 : \mu = \binom{1}{5}, \kappa = 10, \mathbf{\Lambda} = \left(\begin{smallmatrix}6 & 3\\ 3 & 6\end{smallmatrix}\right), \nu = 3$, $M_1 : \mu = \binom{5}{3}, \kappa = 2, \mathbf{\Lambda} = \left(\begin{smallmatrix}2 & 1\\ 1 & 2\end{smallmatrix}\right), \nu = 9$.

Then the Bayes factor

$$\frac{p(\mathbf{x} \mid M_1)}{p(\mathbf{x} \mid M_2)} = \frac{p(\mathbf{y} \mid \mathbf{z}, M_1)}{p(\mathbf{y} \mid \mathbf{z}, M_2)} \frac{p(\mathbf{z} \mid M_1)}{p(\mathbf{z} \mid M_2)}, \tag{3.26}$$

so that if we have set sufficiently broad priors on the nuisance variables to converge to the same parameter estimates in both models, we can bound the second factor on the right hand side, and allow the first factor to dominate. In the next section we show how this idea can be applied to a simple Bayesian network.

3.3.1 Example: Bounds for loosened priors in a simple Bayesian network

In this section we consider how we can obtain bounds for the Bayes factor on identical network structures and likelihoods, but differing choices of prior on certain variables. We consider two networks representing models M_1, M_2. Both networks are assumed to have the same likelihood function, $p(x \mid \theta)$, but different priors, $\pi_i(\theta)$, and so we have

$$p_i(x) = \frac{\pi_i(\theta) p(x \mid \theta)}{p_i(\theta \mid x)}, \tag{3.27}$$

where $p_i(x)$, $p_i(\theta \mid x)$, denote the marginal likelihood and posterior under model M_i. We can therefore express the ratio of the marginal densities under the two models as

$$\frac{p_1(x)}{p_2(x)} = \frac{\pi_1(\theta) p_2(\theta \mid x)}{\pi_2(\theta) p_1(\theta \mid x)}. \tag{3.28}$$

In the case of a large sample of size N, we consider the limiting behaviour of the Bayes factor under different choices of prior as sample size increases. To do this, we consider a situation where the observations of a node X take two values: 0,1 and we have a binomial likelihood with conjugate beta priors $\pi_i(\theta) \sim Be(\alpha_i, \beta_i)$.

It is often proposed to set priors according to a hyperdirichlet distribution (Dawid and Lauritzen [1993], Lauritzen [1996]), corresponding to notional 'counts' in a fictive contingency table. Furthermore it is suggested that the total number of counts is represented by an *equivalent sample size*, which controls the degree of confidence we have in our prior parameter assignment.

Such an assignment may not, however, be appropriate, for example, when the analyst has greater insight into a particular subset of the Bayesian network providing supplementary information. Specific examples could include circumstances where there is a pre-existing model for part of the network interaction, affording a larger effective sample size for this part only, or where additional rules and intervention are known, *a priori* rather than through observation, to apply in particular circumstances.

In addition, as we discussed in the previous section, the analyst may wish to provide weaker priors on the variables of less interest on the basis that she does not wish the Bayes factor to be dominated by poor model performance on these variables.

Using the limiting results from the previous section, we can examine the impact of alternative prior specifications for the Bayesian network. In particular, let $\mathcal{G}$ be a Bayesian network structure of binary variables $\mathbf{X} = \{X_i, \dots X_n\}$. By x_i, pa_{X_i}, we denote specific instances of the random variables, X_i and Pa_{X_i} (the set of parents of X_i) respectively. We assume parameterisations which exhibit both *local* and *global* independence (see, e.g. Koller and Friedman [2009]). In other words, for a general parameterisation $\Theta = \theta_{X_1|Pa_{X_1}}, \dots \theta_{X_n|Pa_{X_n}}$ we assume that, respectively, both

$$p(\Theta) = \prod_i p(\theta_{X_i|Pa_{X_i}}) \tag{3.29}$$

and

$$p(\theta_{X_i|Pa_{X_i}}) = \prod_{\mathbf{u} \in pa_{X_i}} p(\theta_{X_i|\mathbf{u}}). \tag{3.30}$$

In this case, denoting the complete observed data set by D; observations of a subset of variables $\mathbf{U} \subseteq \mathbf{X}$ by $D[\mathbf{U}]$ for any model M, and the subset of $D[\mathbf{U}]$ containing just those observations for which the specific values of a subset of variables $\mathbf{V}$ is equal to an instance $\mathbf{v}$ by $D[\mathbf{U} \mid \mathbf{v}]$, the log likelihood decomposes into a sum of the component conditional likelihoods:

$$\begin{aligned} \log p(D \mid M) &= \sum_i \log p(D[X_i] \mid D[Pa_{X_i}]) \\ &= \sum_i \sum_{\mathbf{u} \in pa_{X_i}} \log p(D[X_i \mid \mathbf{u}] \mid D[Pa_{X_i} \mid \mathbf{u}]). \end{aligned} \tag{3.31}$$

We now compare two alternative parameterisations for $\mathcal{G}$. In model M_1, we set Beta priors on the conditional probabilities in line with a *BDe prior* (see, for example, Heckerman et al. [1995]) with an effective sample size of α and an assumed prior joint distribution p', so that

$$X_i \mid pa_{X_i} \sim Be(\alpha p'(X_i = 1, pa_{X_i}), \alpha p'(X_i = 0, pa_{X_i})). \tag{3.32}$$

In model M_0, we also have Beta priors on conditional probabilities, with the same implied means as for model M_1 but without the constraint of a common effective sample size. Under this model we are free to define

$$X_i \mid pa_{X_i} \sim Be(\alpha_{X_i|pa_{X_i}} p'(X_i = 1, pa_{X_i}), \alpha_{X_i|pa_{X_i}} p'(X_i = 0, pa_{X_i})). \tag{3.33}$$

Combining the decomposition in equation (3.32) with the result from equation (3.25), we have that

$$\log \frac{p(D \mid M_1)}{p(D \mid M_0)} \to \sum_i \sum_{pa_{X_i}} \left(\alpha - \alpha_{X_i|pa_{X_i}}\right) A_i + B_i, \tag{3.34}$$

where

$$A_i = \log \left(\left(\frac{p'(X_i = 1, pa_{X_i})}{p(X_i = 1, pa_{X_i})} \right)^{p'(X_i=1,pa_{X_i})} \left(\frac{1 - p'(X_i = 1, pa_{X_i})}{1 - p(X_i = 1, pa_{X_i})} \right)^{1-p'(X_i=1,pa_{X_i})} \right),$$
$$B_i = -\frac{1}{2} \log \left(\frac{\alpha}{\alpha_{X_i|pa_{X_i}}} \right),$$

and p denotes the true data generating process.

As a special case, where our prior joint probability distribution is uniform across all instances, $p'(x_1, \ldots, x_n) = 2^{-n}$, we have

$$\log \frac{p(D \mid M_1)}{p(D \mid M_0)} \to -\frac{1}{2} \left(\sum_i \sum_{pa_{X_i}} \left(\alpha - \alpha_{X_i|pa_{X_i}}\right) A_i + B_i \right), \tag{3.35}$$

where

$$A_i = \log\left(2p\left(X_i = 1, pa_{X_i}\right)\left(2 - 2p\left(X_i = 1, pa_{X_i}\right)\right)\right),$$
$$B_i = \log\left(\frac{\alpha}{\alpha_{X_i|pa_{X_i}}}\right),$$

and p denotes the true data generating process.

This result allows us to place limiting bounds on the Bayes factor for this network. So by allowing different choices of α, for different variables, as sample size increases we can reduce the impact of the Bayes factor on those aspects of less importance by arranging for the priors on these to be more vague, while allowing the sharper prior choices on the variables of greater interest to have a greater influence on the model selection.

3.3.2 Chapter Summary

In this chapter, we have examined the situation in which we are interested only in a subset of relationships from the full modelled joint distribution. Using the standard Bayes factor will not be optimal in these situations, in that model performance on variables of low interest will contribute to the model assessment.

Instead, we have considered two approaches to tailoring the Bayes factor to this situation. One possibility is to compute a restricted Bayes factor over the variables of interest. By calculating the Bayes factor only in respect of the marginal densities of interest, we can discount the impact of poor performance on nuisance variables.

In certain circumstances we may be able to achieve a similar result by 'loosening' priors on the variables of low interest, so that as models learn through incoming data, the models score similarly on these variables so that their Bayes factor becomes dominated by the performance on the variables of interest. We presented results which allow us to quantify and set bounds on the limiting behaviour of the Bayes factors of variables subject to priors which have been loosened in this way.

Chapter 4

Score based information criteria

4.1 Introduction

We have commented that, in a number of applications, standard Bayes factor model comparison and selection may be inappropriate for decision making under specific, utility-based, criteria. It has been suggested that the use of scoring rules in this context allows greater flexibility: scores can be customised to a client's utility and model selection can proceed on the basis of the highest scoring model.

In this chapter we argue that the approach of comparing the cumulative scores of competing models is not ideal because it tends to ignore a model's ability to 'catch up' through parameter learning. An alternative approach of selecting a model on its maximum posterior score based on a plug in or posterior expected value is problematic in that it uses the data twice in estimation and evaluation.

We therefore introduce a new approach in the form of a context dependent Bayesian information criterion – the *Bayesian Posterior Score Information Criterion (BPSIC)* which is based on a bias corrected posterior predictive expected score which can be tailored to the utility of a model user.

This allows the analyst both to tailor an appropriate scoring function to the needs of the ultimate decision maker and to correct appropriately for bias in using the data on a

posterior basis to revise parameter estimates. We show that this criterion can provide a convenient method of initial model comparison when the number of models under consideration is large or when computational burdens are high. At the end of the chapter, we illustrate the new methods with simulated examples and real data from the UK electricity imbalance market.

4.2 Utility based model selection

We previously outlined that in many applications we are interested in estimating the expected divergence of a model, where the divergence is based on the particular scoring rule which reflects the end user's utility.

Ando [2007] introduced a Bayesian predictive information criterion (BPIC) defined as

$$BPIC = -2E_{\theta|y}\left[\log L(y \mid \theta)\right] + 2n\hat{b}_{\theta}, \tag{4.1}$$

with

$$n\hat{b}_{\theta} = E_{\theta|y}\left[\log(L(y \mid \theta)\pi(\theta))\right] - \log(L(y \mid \hat{\theta}_n)\pi(\hat{\theta}_n)) + \mathrm{Tr}(J_n^{-1}(\hat{\theta}_n)I_n(\hat{\theta}_n)) + p/2, \tag{4.2}$$

where p represents the dimension of the parameter vector θ, $\hat{\theta}_n$ denotes the parameter value which maximises $n^{-1}\log(L(y \mid \theta)\pi(\theta))$ and I_n, J_n are defined as follows:

$$\begin{aligned} I_n(\theta) &= \frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial(\log f_\theta(y_k) + \log\pi(\theta)/n)}{\partial\theta}\frac{\partial(\log f_\theta(y_k) + \log\pi(\theta)/n)}{\partial\theta^T}\right), \\ J_n(\theta) &= -\frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial^2(\log f_\theta(y_k) + \log\pi(\theta)/n)}{\partial\theta\partial\theta^T}\right). \end{aligned} \tag{4.3}$$

Suppose that we observe n observations $y = (y_1, y_2, \ldots y_n)$ and we are considering a candidate model M with probability density $f_\theta := f(y \mid \theta)$, prior density $\pi(\theta)$ and posterior density $\pi(\theta \mid y)$, where we are interested in its ability to minimise the expected divergence induced by the scoring function $S(f, z)$ for future observations z from the true data generating process.

We seek an analogue of the BPIC which allows us to assess, in a Bayesian fashion, the posterior expected quantity $E_z\left[E_{\theta|y}\left[S(f_\theta, z)\right]\right]$, as a measure of discrepancy from the true data generating process for z, for our chosen scoring rule S.

We now generalise the proof in Ando [2007] to establish the following result, where we denote by $\hat{\theta}_n$ the parameter value which maximises $n^{-1}\log(L(y \mid \theta)\pi(\theta))$, and assume that it is unique; we denote the posterior mean by $\bar{\theta}_n$, define the cumulative score $C_S(y \mid \theta) := \sum_{k=1}^{n} S(f_\theta, y_k)$, and define the matrices:

$$
\begin{aligned}
I(\theta) &= E_z\left[\frac{\partial(\log f_\theta(z) + \log \pi_0(\theta))}{\partial\theta}\frac{\partial(\log f_\theta(z) + \log \pi_0(\theta))}{\partial\theta^T}\right], \\
J(\theta) &= -E_z\left[\frac{\partial^2(\log f_\theta(z) + \log \pi_0(\theta))}{\partial\theta\partial\theta^T}\right], \\
J_n^S(\theta) &= -\frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial^2(S(f_\theta, y_k) + \log \pi(\theta)/n)}{\partial\theta\partial\theta^T}\right), \\
U_n^S(\theta) &= \frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial(S(f_\theta, y_k) + \log \pi(\theta)/n)}{\partial\theta}\right).
\end{aligned}
\tag{4.4}
$$

Theorem 6 *Assume that:*

1. $\log \pi_0(\theta) := \lim_{n\to\infty} n^{-1}\log \pi(\theta)$ *exists,*

2. J *is non-singular at* θ_0*, each of the elements of* $I(\theta_0)$ *is continuously differentiable,*

3. *Regularity conditions (see, e.g. Chapter 3 of Barndorff-Nielsen and Cox [1989]) hold that ensure the Laplace approximation of the posterior distribution as* $N(\hat{\theta}_n, n^{-1}J_n(\hat{\theta}_n))$ *is valid,*

4. $U_n^S(\hat{\theta}_n)$ *and* $(\bar{\theta}_n - \hat{\theta}_n)$ *are uncorrelated.*

We define the bias, b_S*, from estimating the posterior expected score by the mean of the posterior scores of the observed data*

$$
b_S := E_y\left[\frac{1}{n}E_{\theta|y}\left[C_S(y \mid \theta)\right] - E_z\left[E_{\theta|y}\left[S(f_\theta, z)\right]\right]\right]. \tag{4.5}
$$

If we estimate the bias, b_S, by the estimator $\hat{b}_S$, where

$$n\hat{b}_S = E_{\theta|y}\left[C_S(y \mid \theta) + \log \pi(\theta)\right] - \left(C_S(y \mid \hat{\theta}_n) + \log \pi(\hat{\theta}_n)\right)$$

$$+\frac{1}{2}\operatorname{Tr}(J_n^S(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n)) + \operatorname{Tr}(J_n^S(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n)I_n(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n))$$

$$-nU_n^S(\hat{\theta}_n)(\bar{\theta}_n - \hat{\theta}_n),$$

then we have

$$(nb_S - n\hat{b}_S) = n\operatorname{cov}(U_n^S(\hat{\theta}_n), \bar{\theta}_n) + O_p(n^{-1/2}). \tag{4.6}$$

4.3 Proof of Theorem 6

We suppose that n observations $y = (y_1, y_2, \ldots y_n)$ are generated from the true process with probability density $g(y)$. We assume that a model M with probability density $f_\theta := f(y \mid \theta)$ and prior density $\pi(\theta)$ is being considered as a candidate model for approximating the true data generating process. We denote the likelihood $L(y \mid \theta) = \prod_{k=1}^n f_\theta(y_k)$. We define $\log \pi_0(\theta) := \lim_{n\to\infty} n^{-1} \log \pi(\theta)$, and in accordance with Assumption 2 of Theorem 6, assume that this exists. We further assume that the parameter vector θ is of dimension p.

Suppose we are interested in the model which minimises the expected divergence induced by the scoring function $S(f, z)$ for observations z from the true data generating process. We define the cumulative score $C_S(y \mid \theta) := \sum_{k=1}^n S(f_\theta, y_k)$.

We denote θ_0, $\hat{\theta}_n$ as the parameter values which maximise $E_z\left[\log(f_\theta(z)\pi_0(\theta))\right]$ and $n^{-1}\log(L(y \mid \theta)\pi(\theta))$ respectively, and assume that these are unique. We define $\bar{\theta}_n$ as the posterior mean for θ, and define the matrices and estimators $I(\theta)$, $J(\theta)$, $J^S(\theta)$, $U^S(\theta)$, $I_n(\theta)$, $J_n(\theta)$, $J_n^S(\theta)$, $U_n^S(\theta)$ as in Equations 4.3 and 4.4.

Ando [2007] establishes the following results for θ_n and θ_0 assuming the appropriate regularity conditions in Assumption 3 of Theorem 6.

Lemma 7 *$(\hat{\theta}_n - \theta_0)$ converges in distribution to $N(0, n^{-1}J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0))$.*

Proof. See Ando [2007] ■

Lemma 8 *In accordance with Assumption 4 of Theorem 6, if we assume that appropriate regularity conditions (see, e.g. Chapter 3 of Barndorff-Nielsen and Cox [1989]) hold that ensure the Laplace approximation of the posterior distribution as* $N(\hat{\theta}_n, n^{-1}J_n(\hat{\theta}_n))$ *is valid, we have:*

$$E_y\left[E_{\theta|y}\left[(\theta-\theta_0)(\theta-\theta_0)^T\right]\right] = \frac{1}{n}J^{-1}(\theta_0) + \frac{1}{n}J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0) + O_p(n^{-3/2}). \quad (4.7)$$

Proof. See Ando [2007] ■

Proof of Theorem 6. The proof uses the method which is introduced in Ando [2007], but adjusted to allow for the fact that the estimator $\hat{\theta}_n^\star$ which maximises the scoring function may differ from the posterior mode $\hat{\theta}_n$. In particular, this means that the expansion around the posterior mode also requires the inclusion of the relevant first derivatives. We express the bias as the sum of three expected values:

$$\begin{aligned} E_1 &= E_y\left[\frac{1}{n}E_{\theta|y}\left[C_S(y \mid \theta)\right] - \frac{1}{n}(C_S(y \mid \theta_0) + \log\pi(\theta_0))\right], \\ E_2 &= E_y\left[\frac{1}{n}(C_S(y \mid \theta_0) + \log\pi(\theta_0)) - E_z\left[S(f_{\theta_0}, z) + \log\pi_0(\theta_0)\right]\right], \\ E_3 &= E_y\left[E_z\left[S(f_{\theta_0}, z) + \log\pi_0(\theta_0)\right] - E_z\left[E_{\theta|y}\left[S(f_\theta, z)\right]\right]\right]. \end{aligned} \quad (4.8)$$

Approximating E_1

To approximate E_1, we perform a Taylor expansion of $C_S(y \mid \theta_0) + \log\pi(\theta_0)$ around the posterior mode $\hat{\theta}_n$, where we obtain:

$$C_S(y \mid \theta_0) + \log\pi(\theta_0) = C_S(y \mid \hat{\theta}_n) + \log\pi(\hat{\theta}_n) + n(\theta_0 - \hat{\theta}_n)U_n^S(\hat{\theta}_n) - n/2(\theta_0 - \hat{\theta}_n)^T J_n^S(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n) + O_p(n^{-1/2}),$$

and so we have

$$E_1 = \frac{1}{n}E_y\left[E_{\theta|y}\left[C_S(y \mid \theta)\right] - (C_S(y \mid \hat{\theta}_n) + \log\pi(\hat{\theta}_n))\right]$$

$$-E_y\left[U_n^S(\hat{\theta}_n)(\theta_0-\hat{\theta}_n)\right]+\frac{1}{2}\operatorname{Tr}(E_y\left[J_n^S(\hat{\theta}_n)(\theta_0-\hat{\theta}_n)(\theta_0-\hat{\theta}_n)^T\right])+O_p(n^{-3/2}). \quad (4.9)$$

Using Lemma 7, we then have

$$E_1=\frac{1}{n}E_y\left[E_{\theta|y}\left[C_S(y\mid\theta)\right]-(C_S(y\mid\hat{\theta}_n)+\log\pi(\hat{\theta}_n))\right] \quad (4.10)$$

$$+\frac{1}{2n}\operatorname{Tr}(J^S(\theta_0)J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0))-E_y\left[U_n^S(\hat{\theta}_n)(\theta_0-\hat{\theta}_n)\right]+O_p(n^{-3/2}).$$

Approximating E_2

We can ignore the term E_2 as we have:

$$E_2=E_y\left[S(f_{\theta_0},y)+\log\pi_0(\theta_0)\right]-E_z\left[S(f_{\theta_0},z)+\log\pi_0(\theta_0)\right]-\log\pi_0(\theta_0)+\frac{1}{n}\log\pi(\theta_0), \quad (4.11)$$

and using Assumption 1 of Theorem 6 that $\log\pi(\theta)=O(1)$, we have

$$E_2=o_p(n^{-1}). \quad (4.12)$$

Approximating E_3

For the term E_3, we perform a Taylor expansion around θ_0. Writing

$$E_3=E_z\left[S(f_{\theta_0},z)+\log\pi_0(\theta_0)\right]-E_y\left[E_{\theta|y}\left[E_z\left[(S(f_\theta,z)+\log\pi_0(\theta)\right]\right]\right]$$

$$+E_y\left[E_{\theta|y}\left[\log\pi_0(\theta)\right]\right], \quad (4.13)$$

if we expand around θ_0, then we have

$$E_3=E_z\left[S(f_{\theta_0},z)+\log\pi_0(\theta_0)\right]-E_z\left[(S(f_{\theta_0},z)+\log\pi_0(\theta_0)\right]-U^S(\theta_0)E_y\left[E_{\theta|y}\left[(\theta-\theta_0)\right]\right]$$

$$+\frac{1}{2}\operatorname{Tr}(J^S(\theta_0)E_y\left[E_{\theta|y}\left[(\theta-\theta_0)(\theta-\theta_0)^T\right]\right])+E_y\left[E_{\theta|y}\left[\log\pi_0(\theta)\right]\right]+O_p(n^{-3/2}). \quad (4.14)$$

Applying Lemma 8 gives

$$E_3 = \frac{1}{2n}\operatorname{Tr}(J^S(\theta_0)(J^{-1}(\theta_0) + J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0)) + \frac{1}{n}E_y\left[E_{\theta|y}\left[\log\pi(\theta)\right]\right]$$

$$-U^S(\theta_0)E_y\left[E_{\theta|y}\left[(\theta-\theta_0)\right]\right] + O_p(n^{-3/2}). \tag{4.15}$$

Approximating total bias

Combining the terms then gives the bias estimator

$$nb_S = E_y\left[E_{\theta|y}\left[C_S(y\mid\theta) + \log\pi(\theta)\right] - \left(C_S(y\mid\hat{\theta}_n) + \log\pi(\hat{\theta}_n)\right)\right]$$

$$+\frac{1}{2}\operatorname{Tr}(J^S(\theta_0)J^{-1}(\theta_0)) + \operatorname{Tr}(J^S(\theta_0)J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0))$$

$$-nE_y\left[U_n^S(\hat{\theta}_n)(\theta_0-\hat{\theta}_n)\right] - nU^S(\theta_0)E_y\left[E_{\theta|y}\left[(\theta-\theta_0)\right]\right] + O_p(n^{-1/2}). \tag{4.16}$$

After some rearrangement, the final two terms can be written as

$$-nE_y\left[U_n^S(\hat{\theta}_n)(\theta_0-\hat{\theta}_n) + U^S(\theta_0)(\bar{\theta}_n-\theta_0)\right], \tag{4.17}$$

which, in turn, directly from our definitions

$$= -nE_y\left[U_n^S(\hat{\theta}_n)(\bar{\theta}_n-\hat{\theta}_n)\right] + nE_y\left[(U_n^S(\hat{\theta}_n) - U^S(\theta_0))(\bar{\theta}_n-\theta_0)\right]. \tag{4.18}$$

The first term vanishes where posterior modes and means are equal (as would be the case, for example, under conjugate symmetric priors). Where not, we make use of Assumption 5 in Theorem 6 that $U_n^S(\hat{\theta}_n)$ and $(\bar{\theta}_n - \hat{\theta}_n)$ are uncorrelated, to express the expectation as $-nU_n^S(\hat{\theta}_n)(\bar{\theta}_n-\hat{\theta}_n) + O_p(n^{-1/2})$.

We next arrange the second term as

$$E_y\left[\sqrt{n}(U_n^S(\hat{\theta}_n) - U^S(\theta_0))\sqrt{n}(\bar{\theta}_n-\theta_0)\right]$$

$$= E_y\left[\sqrt{n}(U_n^S(\hat{\theta}_n) - U^S(\theta_0))\right]E_y\left[\sqrt{n}(\bar{\theta}_n-\theta_0)\right] + n\operatorname{cov}(U_n^S(\hat{\theta}_n), \bar{\theta}_n)$$

$$= n \operatorname{cov}(U_n^S(\hat{\theta}_n), \bar{\theta}_n) + O_p(n^{-1/2}). \tag{4.19}$$

Replacing quantities with their estimators, we then have:

$$nb_S = E_{\theta|y}\left[C_S(y \mid \theta) + \log \pi(\theta)\right] - (C_S(y \mid \hat{\theta}_n) + \log \pi(\hat{\theta}_n)) - nU_n^S(\hat{\theta}_n)(\bar{\theta}_n - \hat{\theta}_n)$$

$$+\frac{1}{2}\operatorname{Tr}(J_n^S(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n)) + \operatorname{Tr}(J_n^S(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n)I_n(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n))$$

$$+n \operatorname{cov}(U_n^S(\hat{\theta}_n), \bar{\theta}_n) + O_p(n^{-1/2}), \tag{4.20}$$

from which the result follows. ■

We now introduce the further assumption that the covariance term on the right hand side of Equation 4.6 is small in comparison to $n\hat{b}_S$. In practice, we have found common scoring rules satisfy this condition. For example, for the logarithmic scoring rule the term is zero. In the case of piecewise linear scoring rules used for quantile scoring (which we consider later) these have values of the derivative $U_n^S(\hat{\theta}_n)$ dominated by a constant term which depends solely on the number of observations exceeding a quantile estimate, and therefore have an extremely low covariance with the level of the parameter estimate $\hat{\theta}_n$. Where novel scoring rules are being considered, possible violation of this condition can be investigated, for example by simulation.

In the light of Theorem 6 and the comment above, we therefore propose a **Bayesian Posterior Score Information Criterion** which is defined as:

$$BPSIC = -2E_{\theta|y}\left[C_S(y \mid \theta)\right] + 2n\hat{b}_S. \tag{4.21}$$

We follow other more conventional information criteria in the choice of sign and models with lower values of the BPSIC are therefore preferred.

Note that:

1. When the selected scoring rule, S, is equal to the logarithmic score, then BPSIC corresponds to BPIC.

2. All the relevant quantities can be readily computed, for example from a MCMC posterior sample. This enables its calculation to be incorporated as a standard routine in the initial evaluation of multiple models, without the need to perform the multiple estimation runs required for cross-validation.

3. The computation of $J^S(\hat{\theta}_n)$ requires that the score function should have a finite second derivative. This may not be the case at all points for certain scoring functions, for example, absolute loss, or piecewise linear functions. In practice, we have found that this tends not to be problematic because the points at which the derivatives of the scoring rule do not exist will tend not to concentrate around the posterior mode. However, we indicate in the examples in the next section how routine modifications can ensure problems do not occur.

Use of the BPSIC should readily facilitate an initial comparison of future expected utilities of models. In addition, if, for example, MCMC output is stored, then a future user of a model should be able to re-assess its performance based on an alternative scoring rule with a fairly straightforward recalculation.

In the next section we illustrate the performance of the BPSIC with three examples based on stylised simulated data. Then in Section 4.5 we illustrate the application to the problem of predicting quantiles of UK electricity imbalance prices.

4.4 Simulation examples

4.4.1 Performance on different score functions

In this example we compare estimates obtained using BPSIC to the actual bias. We consider different score functions and situations in which the model is correctly and incorrectly specified. This gives an insight into the performance of the BPSIC approximation in a variety of applications.

In the correctly specified scenario, we consider a normal model M_1 with unknown mean μ, known variance $\sigma^2 = 0.5^2$, and a conjugate prior $\mu \sim N(0, 0.1^2)$. We assume that the

true data generating process is normally distributed with mean 0 and variance 0.5^2. In the incorrectly specified scenario, the normal model M_1 has unknown mean μ, known variance $\sigma^2 = 0.5^2$, and a conjugate prior $\mu \sim N(0, 5^2)$, and we assume that the true data generating process is normally distributed with mean 1 and variance 2^2.

We consider four score functions (where we define these as the negative of the corresponding loss function), where the score $S(f_\theta, y_k)$ results under the model expressed by the density function f at the parameter value θ, if the value y_k is observed.

- Logarithmic predictive density. We define $S(f_\theta, y_k) = \log f_\theta(y_k)$. As we have seen, maximising this score corresponds to minimising the Kullback-Leibler divergence.

- Quadratic score. We define $S(f_\theta, y_k) = -(\mu(f_\theta) - y_k)^2$, where $\mu(f_\theta)$ denotes the predictive mean of the distribution with density f. Although for the purposes of the normal model example here, this reduces to a scaled version of the log density when a vague prior is chosen, when we include the more informative prior specification we have adopted under the correctly specified model scenario, it also results in a different weighting between prior and score function.

- Absolute loss. We define $S(f_\theta, y_k) = -\left|\mu(f_\theta) - y_k\right|$. We remark that there are undefined second derivatives at $\mu(f_\theta) = y_k$. Any problems encountered can be addressed by approximating this by the Huber loss function, defined as

$$S(f_\theta, y_k) = \begin{cases} -(\mu(f_\theta) - y_k)^2/2, & \text{if } |\mu(f_\theta) - y_k| < k \\ k(|\mu(f_\theta) - y_k| - k/2) & \text{otherwise.} \end{cases}$$

- Quantile loss. This time, our focus of interest is in being able to forecast a specific quantile – perhaps for a risk management application. We select a quantile scoring rule, reflecting a focus on our ability to forecast the 0.95 quantile. A number of quantile scoring rules have been established (see Gneiting and Raftery [2007]); here we make use of the *asymmetric piecewise linear scoring function* (see Gneiting

[2011] defined by

$$S(f_\theta, y_k) = (y_k - \tau(f_\theta))(\mathbb{1}(\tau(f_\theta) > y_k) - \tau)$$

where $0 < \tau < 1$ is the quantile of interest, and $\tau(f_\theta)$ denotes the value predicted by the density f_θ.

As with absolute loss, one common feature of these quantile scoring rules is that they are piecewise linear, therefore having undefined second derivatives for some values. For the computation of the BPSIC it is possible to make an adjustment by making use of the *quantile Huber loss* proposed by Aravkin et al. [2014], which takes the form:

$$\rho_\tau(f_\theta, y_k) = \begin{cases} \tau \left| y_k - \tau(f_\theta) \right| - \dfrac{\kappa\tau^2}{2}, & \text{if } y_k - \tau(f_\theta) < -\tau\kappa \\ (1-\tau) \left| y_k - \tau(f_\theta) \right| - \dfrac{\kappa(1-\tau)^2}{2}, & \text{if } y_k - \tau(f_\theta) > (1-\tau)\kappa \\ \dfrac{1}{2\kappa}(y_k - \tau(f_\theta))^2 & \text{otherwise,} \end{cases}$$

where the value of κ is selected by the user as the threshold within which a quadratic approximation replaces the corresponding piecewise linear scoring rule.

The graphs below show the result of comparing the average BPSIC bias with the average actual bias (based on simulating future observations from the true distribution). Figure 4.1 shows the results in the correctly specified model case; Figure 4.2 illustrates the incorrectly specified case.

We observe that the bias in the first case where the prior is more informative is lower, reflecting the greater weighting given to the prior compared to the new data. The scale of the bias is dominated by the natural scale of the scores themselves. We conjecture that there is an additional effect in that greater bias is likely to be seen when we use scores which are 'closer' to the logarithmic score: asymptotically, this score will be maximised under Bayesian updating. We comment on this in our conclusion.

To illustrate the impact of employing different score functions on model selection, we

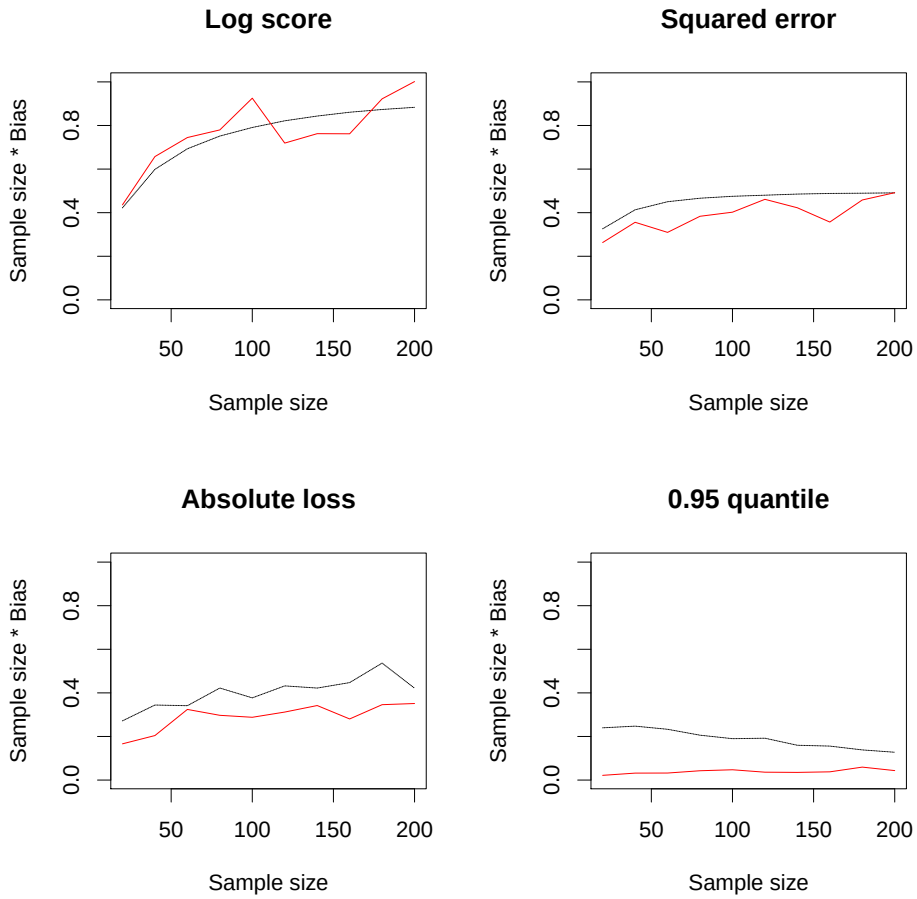

Figure 4.1: Performance of actual bias compared to asymptotic (BPSIC) bias. The true data generating process is given by a $N(0, 0.5^2)$ distribution. The model being assessed, M_1, is a normal distribution with unknown mean and known variance equal to the true variance. The mean has a conjugate prior $\mu \sim N(0, 0.1^2)$. The figure shows the simulated average actual bias (7,000 simulations, with computation of the relevant expectations for each simulation computed by 1,000 posterior parameter simulations), shown by the solid red line and the average asymptotic bias (dotted black line) under four different loss functions.

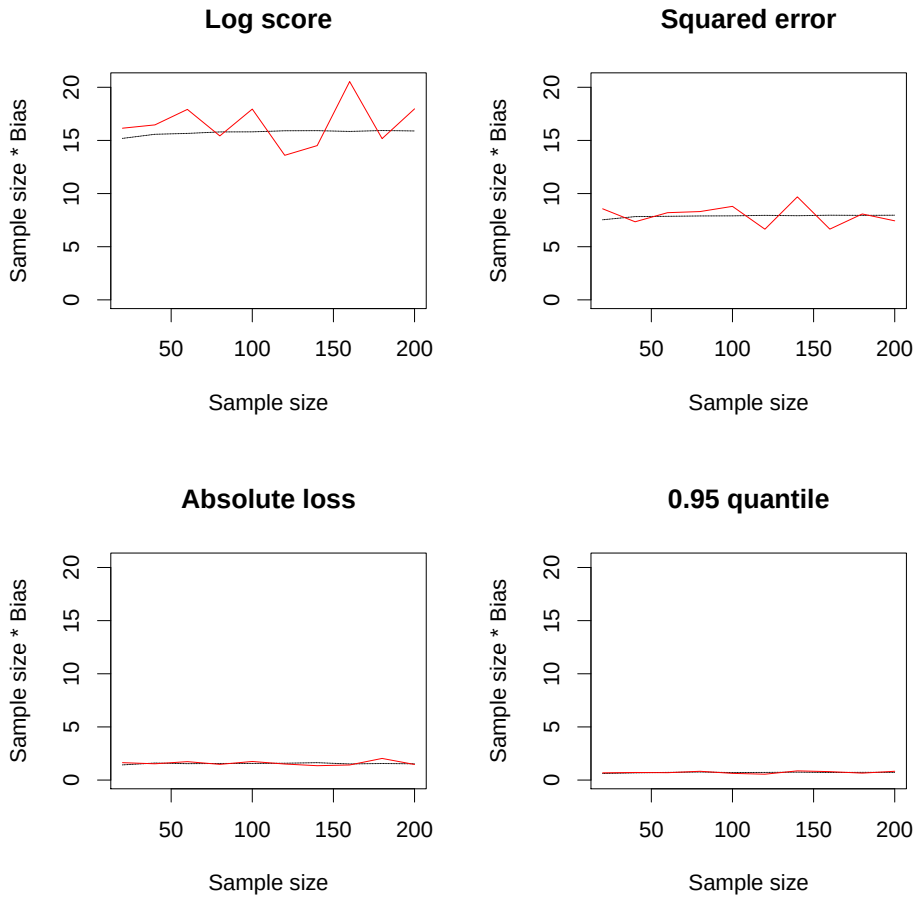

Figure 4.2: Performance of actual bias compared to asymptotic (BPSIC) bias - misspecified model. The true data generating process is given by a $N(1, 2^2)$ distribution. The model being assessed, M_1, is a normal distribution with unknown mean and fixed variance of 0.5^2. The mean has a conjugate prior $\mu \sim N(0, 5^2)$. The figure shows the simulated average actual bias (7,000 simulations, with computation of the relevant expectations for each simulation computed by 1,000 posterior parameter simulations), shown by the solid red line and the average asymptotic bias (dotted black line) under four different loss functions.

consider a stylised example. 200 data points were simulated from a mixture model consisting of a Weibull(Shape = 5, Scale = 20) and a lognormal (Log mean = 3, Log SD = 0.8) distribution, where the mixing proportion was weighted 0.8 Weibull and 0.2 lognormal.

MCMC was used to estimate two candidate models: a Weibull model and a lognormal model. For both models, Uniform(0, 100) distributions were used as parameter priors. Figure 4.3 below shows the fitted posterior predictive densities together with the empirical data density.

Table 4.1 shows the BPSIC values under each of the four measures considered earlier in this section, for each of the two models.

Table 4.1: Comparison of BPSIC for different scoring functions

	Log Score	Squared error	Absolute Loss	0.95 Quantile
Weibull model	1,549.5	45,479.7	227.8	669.0
Lognormal model	1,502.3	46,890.5	219.3	759.6

Note that the lognormal model is favoured under the logarithmic score and absolute loss, whereas the Weibull model is favoured if the squared error loss or quantile score is used. If our interest is in the tail quantiles of the distribution, the Weibull model may be a more appropriate model choice, even though it would not be chosen through a default log score procedure. We return to this theme in Section 4.5, when we consider a problem of model choice motivated by a risk management requirement where the appropriate utility relates to forecasts of distribution quantiles.

4.4.2 Comparison with cross-validation

The previous example showed a reasonable fit between the estimates provided by BPSIC and the average bias across a number of loss functions. However, the practical application of the criterion will depend also on the amount of additional variance introduced through the bias correction.

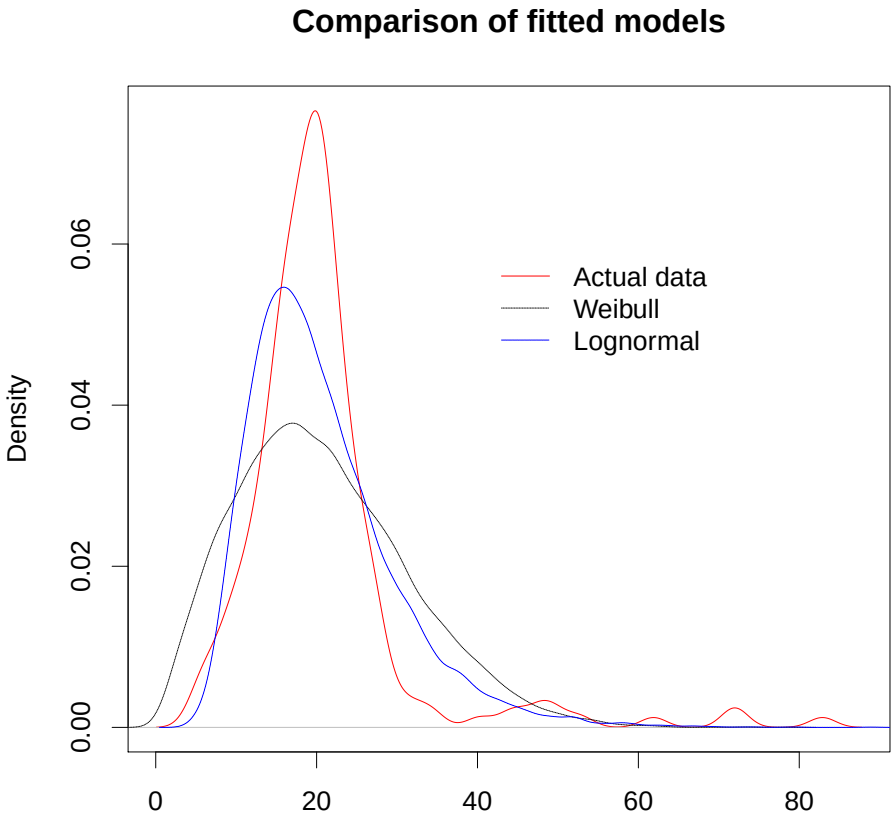

Figure 4.3: Comparison of posterior predictive densities relating to two candidate models estimated using MCMC. The observed data consists of 200 points sampled from a mixture of a Weibull(Shape = 5, Scale = 20) and a lognormal (Log mean = 3, Log SD = 0.8) distribution, where the mixing proportion was weighted 0.8 Weibull and 0.2 lognormal. In the text we illustrate that under different scoring functions within the BPSIC, different candidate models will be selected.

In our next example we compare the performance of the BPSIC and leave one out cross validation (LOO-CV). As before, we use the same mis-specified example, with the model M_1 having unknown mean μ, known variance $\sigma^2 = 0.5^2$, and a conjugate prior $\mu \sim N(0, 5^2)$, and the true data generating process being normally distributed with mean 1 and variance 2^2.

This time, however, we select the quantile scoring rule reflecting a focus on model performance on forecasting the 0.95 quantile. We simulate 2,000 scenarios in which a sample of 100 observations is used to generate a LOO-CV score, an (unadjusted) posterior score, the BPSIC, and simulated 'true score'. The results are shown in Figure 4.4.

The LOO-CV and BPSIC estimation errors are extremely close, with the unadjusted posterior score positively biased. The standard deviation of errors is almost identical in the two methods. The amount of the bias correction is shown in the second graph, and we also show the standard deviations of the individual 'one left out' predictive scores which are averaged to form the LOO-CV estimate.

Table 4.2 shows the computation time on a medium specification PC for the calculation of the LOO-CV scores and the BPSIC, using the example above, but with different sample sizes, n.

Table 4.2: Comparison of computation times (minutes) for BPSIC and LOO-CV

	$n = 100$	$n = 200$	$n = 400$	$n = 1000$
LOO-CV	10.93	21.93	43.72	110.03
BPSIC	9.02	17.40	34.71	80.42

In this example, there is a modest saving in computation time using BPSIC compared to LOO-CV. However the model is a trivial one to re-estimate within each cross-validation. For more complex models, where the Bayesian updating is more time consuming and might involve, for example, MCMC, LOO-CV will incur an additional overhead approximately equal to the sample size multiplied by the additional time necessary to estimate the posterior parameters, compared to a single parameter estimation step for

the BPSIC. Therefore, in situations where this estimation step dominates execution time, LOO-CV could become prohibitively costly.

In the situation in which LOO-CV estimates are expensive to obtain, we might be tempted to undertake a randomised selection of a subset of samples. However, in this example, the additional variability introduced by using a smaller LOO sub-sample is significantly in excess of that introduced by the bias adjustment by the BPSIC.

We should remark that one advantage of the cross-validated score over the BPSIC is that it would enable us to better assess performance of the predictive density of the *posterior predictive score*. For many prediction problems, we are more concerned with how the posterior predictive score will perform than we are with the average of the scores across the posterior distribution. Note that, as the BPSIC is defined as an average divergence (and therefore an estimate of $E_z\left[E_{\theta|y}\left[S(f_\theta, z)\right]\right]$), this means that for concave score functions such as the log score, we will have

$$E_z\left[E_{\theta|y}\left[S(f_\theta, z)\right]\right] \leq E_z\left[S(E_{\theta|y}\left[f_\theta\right], z)\right], \tag{4.22}$$

where $E_{\theta|y}\left[f_\theta\right]$ is the posterior predictive density. In this example, this will mean that if our concern is *when* we should employ the posterior predictive distribution, this will typically be earlier than that suggested by the BPSIC.

4.4.3 Posterior averaged performance in terms of 'catching up'

We commented previously that the practice of comparing models on the basis of their cumulative scores may be less than optimal. Informally, if we are interested in making use of the models to make future predictions, we may be less concerned about their early performance than in their more recent 'track record'. This is likely to be particularly pertinent in a high dimensional setting, where we require increasingly large 'training sets' to calibrate model parameters. van Erven et al. [2012] study this 'catch up' effect, and propose a solution in which a prior is placed over a switching distribution, governing which model should be used in making predictions at a given point in time.

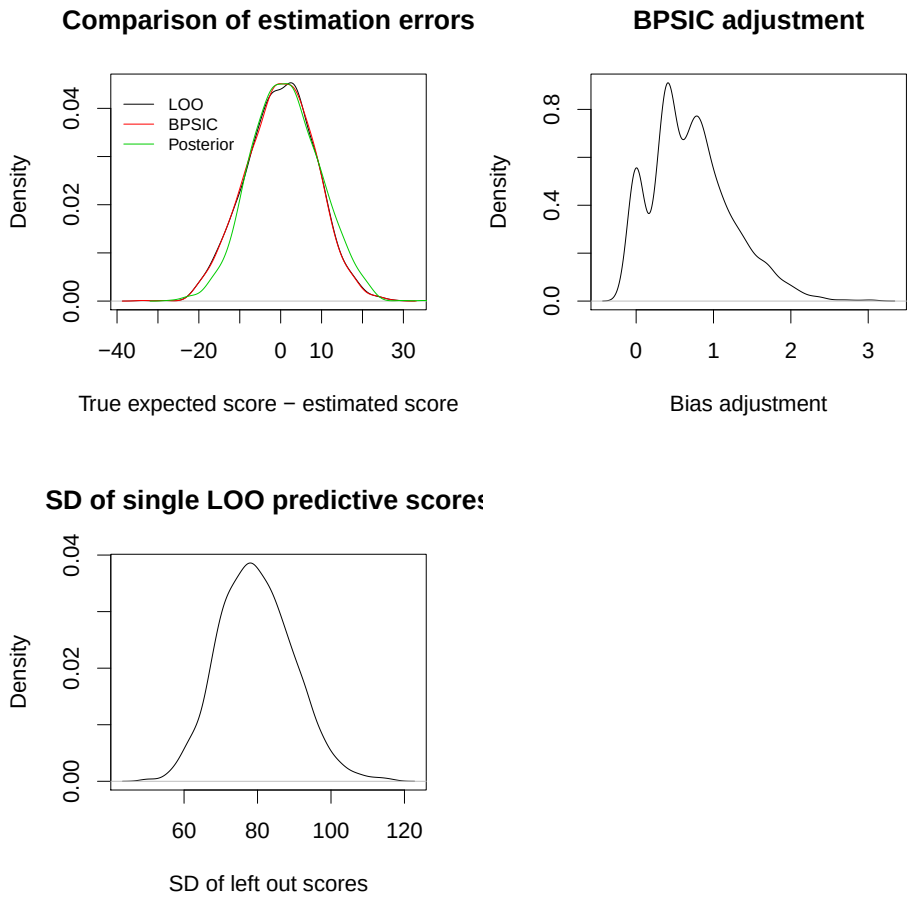

Figure 4.4: Comparison of the BPSIC with leave one out cross validation. The true data generating process is given by a $N(1, 2^2)$ distribution. The model being assessed, M_1, is a normal distribution with unknown mean and fixed variance of 0.5^2. The mean has a conjugate prior $\mu \sim N(0, 5^2)$. The scoring function chosen is the 0.95 asymmetric piecewise linear quantile loss function. The figure shows the results of 2,000 simulations of a sample size of 100. In each simulation the BPSIC is calculated using a sample of 1,000 from the posterior distribution, and the 'true' scores are calculated on a new sample of 1,000 observations generated from the true distribution. The first graph shows the estimation error resulting from the leave one out, unadjusted posterior and adjusted (BPSIC) score. The second graph shows the variation in the bias adjustment for the BPSIC, and the final graph shows the standard deviation of the individual 'left out' scores which are averaged to form the LOO-CV estimate.

An alternative approach of 'calibrating' the models to a similar level of information on an initial training sample, and then comparing models on their subsequent performance has been proposed in Xu et al. [2011]. Both approaches retain the Bayes factor (or equivalently, the cumulative log score) as the selection metric but, instead, make adjustments to compensate for the catch up effect itself.

We suggest that an alternative approach is to discard the Bayes factor altogether as being inappropriate for this type of problem. Instead we would plan to estimate the expected **future** utility. Here we examine the ability of the BPSIC to assess the model's performance based on its **current state** (that is, taking into account parameter learning) as an alternative to using modified Bayes factor selection. We use log predictive utility here. However, we note that our analysis below could be repeated using other utilities.

Accordingly, we suppose that the true model is normally distributed $N(0.2, 1^2)$. We wish to compare two models M_1: a fixed model normally distributed $N(0, 1^2)$ and model M_2 with known variance 1^2 but unknown mean $\mu \sim N(1, 4^2)$. The relatively vague prior on μ in model M_2 means that, assessed on a sequential basis as data is received, M_1 will initially perform better. However, after sufficient observations, model M_2 will become the preferred model.

Figure 4.5 shows the results based on 1,000 simulations. If the Bayes factor (equivalently cumulative log score) is used then on average, model M_2 will only be chosen when the cumulative log score difference is lower than 0, that is after approximately 170 observations. Naive assessment based on the uncorrected posterior log score (that is, the posterior Bayes factor of Aitkin [1991]) will result in selecting model M_2 immediately. If the true posterior log score is used (with knowledge of the true data generating process) then model M_2 should be preferred on average much earlier – after approximately 50 observations. We obtain a very similar result using the corrected estimate from the BPSIC (bottom left hand graph). Of course, we would also obtain very similar results using cross validation, however, as we have already commented, for a large class of models, this might be expected to be much more computationally intensive.

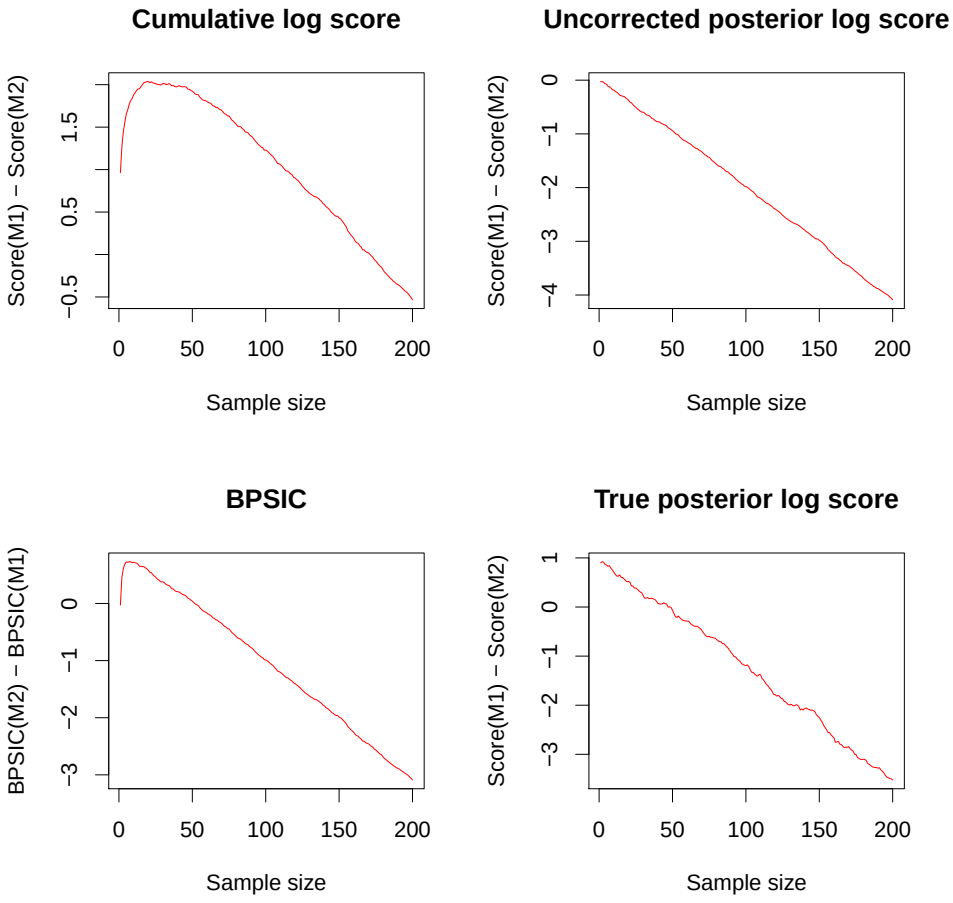

Figure 4.5: Comparison of the performance of the cumulative log score and posterior average (estimated and true) log scores. The true data generating process is given by a $N(0.2, 1^2)$ distribution. We compare two models: M_1 is a fixed model consisting of a normal distribution $N(0, 1^2)$. Model M_2 is a model with known (true) variance and unknown mean μ, $N(\mu, 1^2)$, where the mean has a conjugate prior $\mu \sim N(1, 4^2)$. The figure shows the comparison scores averaged over 1,000 simulations.

4.5 Quantile prediction - UK electricity market imbalance

Where the end goal is to select a model which is used to provide an estimate of 'tail risk', the use of a suitably bias corrected scoring rule can be particularly appropriate. In this section, we apply the BPSIC to the problem of risk management of imbalance exposures to UK electricity market participants. Within the UK, an electricity balancing mechanism is managed by the System Operator (National Grid) to ensure security of supply. Market participants are required to inform the System Operator of their forecast output (in the case of generators) and demand (in the case of suppliers) approximately one hour in advance of each half hour's electricity production.

Typically, it will be necessary for the System Operator to intervene to ensure the actual electricity generated in a given period meets actual demand (which will be different from that implied by the aggregate of forecasts received due to forecast error). In the situation we consider here, the overall system is *short*, in other words it is necessary for the System Operator to seek additional sources of generation (for example, requesting additional short term generation be activated or requiring certain high users to reduce demand). The cost of these activities is reflected in the *System Buy Price* (SBP) charged to those who have under-forecast demand or over-forecast supply, and this will typically be significantly higher than the prevailing market price. In this situation, the *System Sell Price* (SSP) will reflect a prevailing market price.

Accurate quantification of the amount of risk exposure to imbalance volumes is an important consideration for all market participants. Typically, participants might agree a core pricing model with fixed parameter values which is justified with reference to overall fit to imbalance prices. In addition, suppliers of energy will need to charge a risk premium on all contracts which contribute to system imbalance. The risk premium is often based on an assessment of the 95th percentile, in accordance with market risk practice within the financial services industry. For the quantification of the additional risk premium it might be necessary to justify to the regulator if a different model to the core model would be more appropriate to capture these aspects.

Imbalance data for each day in the period 17th October 2011 to 28th May 2014 (the full length of the historical period stored) was obtained from the Elexon data portal (https://www.elexonportal.co.uk). The dataset used for modelling comprised the Net Imbalance Volume (NIV) (the total amount of electricity (in MWh) by which the system was short or long compared to forecast demand), SBP and SSP (both denominated in GBP/MWh). Prices and balancing behaviour vary throughout the day depending on the degree of demand across the day, and for this exercise we used data only for Settlement Period 16 (this corresponds to a particular half hourly generation period on each day between 7.30 am and 8.00 am in the winter and between 6.30 am and 7.00 am in the summer). This resulted in a sample size of 438.

We would expect the SBP to become more stressed in periods where the NIV is higher. This might be the case, for example, where a power station suffers an unforeseen outage, as in these situations the System Operator will be required to procure a substantial amount of energy at very short notice, often being forced to make use of extremely high cost sources of generation and/or high bids from commercial generators. The degree of stress can be measured by the ratio of the SBP to a measure of typical prices which are prevailing in the market - for this purpose we use the ratio of SBP to SSP.

Panels a) and b) in Figure 4.6 show a plot of the SBP/SSP compared to the NIV. As can be seen, in addition to a positive relationship between the NIV and the SBP/SSP ratio there is also a significant skew in the residuals, as we would expect from an increasingly expensive 'supply stack' of generation.

The skew-normal distribution (Azzalini [1986]) has been used successfully to reflect skewness without the need for ancillary data transformation and was used to model this aspect of the data. In particular, we selected a linear regression model with skew-normal residuals of the form

$$f_{SN}(y_i;\beta_1,\beta_2,\omega^2,\alpha) = \frac{2}{\omega}\phi\left(\frac{y_i-(\beta_1+\beta_2 x_i)}{\omega}\right)\Phi(\alpha\omega^{-1}(y_i-(\beta_1+\beta_2 x_i))). \qquad (4.23)$$

MCMC was used to obtain posterior estimates (see Fruhwirth-Schnatter and Pyne

[2010]). In the first model, we fitted simultaneously all parameters: the linear regression slope and intercept terms, together with the skewnormal parameters, using a vague normal gamma prior (a diagonal matrix with entries of 0.01 for the precision matrix, a mean of zero and $Gamma$(Shape = 0.01, Rate = 0.01) distribution). The posterior distributions from 12,000 simulations with a burn in of 4.000 are shown in Figure 4.6.

In practice, participants in the market might make use of alternative models for different purposes. For example, models used for calculating a 'cost of risk' might use one set of assumptions, and models used to calculate expected levels of imbalance loss might use another set. This might need to be justified to a regulator to ensure that parameters were not being 'tuned' to benefit a particular market participant, and that they reflected an objectively justifiable aspect of model performance. For example, if a different model was being used for market pricing from that used for general market analysis purposes, there would need to be a clear justification of what features of model performance made the models applicable for their areas of use.

To illustrate this aspect, as alternative models, which would need to be separately justified to a regulator, we constrained the intercept term, β_1 at different values between 0.2 and 1.2. BPSIC values were computed for the standard logarithmic score and also the 0.95 quantile scoring rule, reflecting the possible focus of interest of a risk management decision using this model to price an appropriate risk premium. In Figure 4.7, we show the corresponding BPSIC at various values. In particular, the model with the value of β_1 close to that fitted freely gives the highest BPSIC logarithmic score. However, if our interest is in fitting accurately to the prediction of the 0.95 quantile, the graphs show that models with a lower intercept value provide a greater expected future utility.

Table 4.3 summarises the information criteria output for the fitted model, together with the highest scoring models under logarithmic and quantile scoring criteria. Although not considered here, it is easy to see how such a table could be extended to include other scoring rules reflecting the diverse utilities of the possible future user base for a model. Such an extension could enable a more informed selection of the most appropriate model implementation and parameterisation for a particular need.

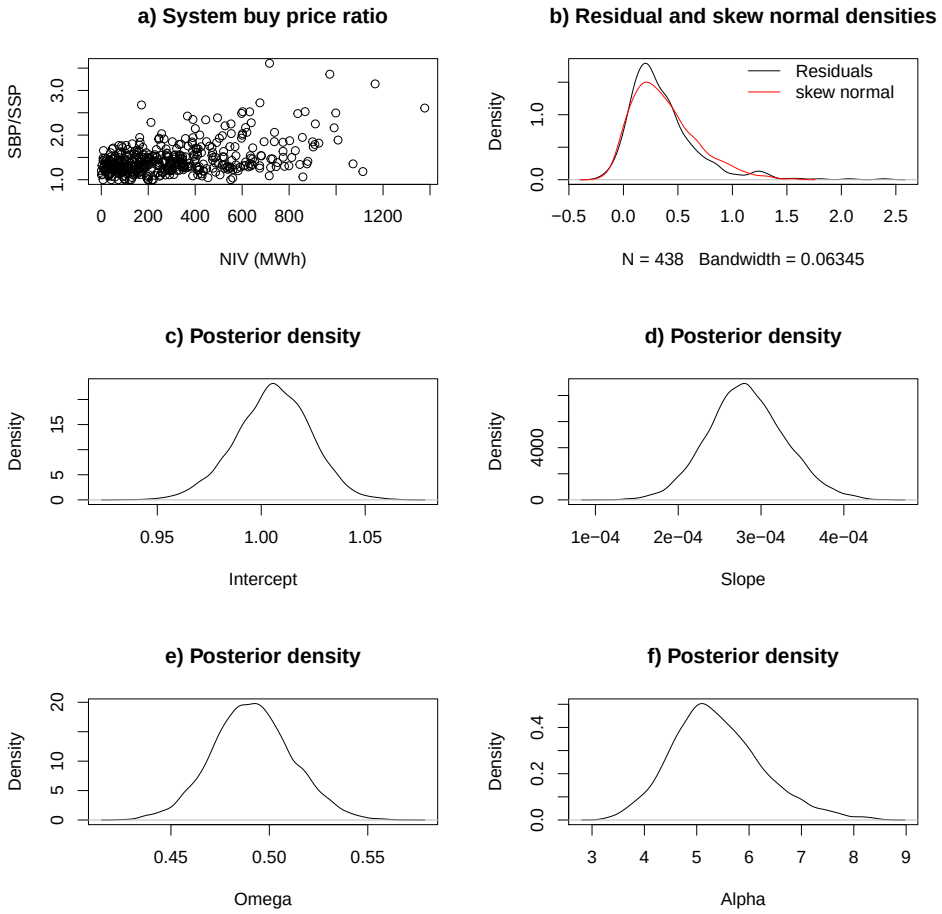

Figure 4.6: Estimation of the relationship between Net Imbalance Volume (NIV) and ratio of System Buy Price to System Sell Price (SBP/SSP). Panel a) shows data between October 2011 and May 2014 relating to settlement period 16 for those occasions when the system was short. MCMC was used to estimate jointly the linear regression parameters relating the SBP/SSP ratio to the NIV. In order to compare visually the residual distribution against the skew normal distribution, Panel (b) shows residuals from the linear regression model (see equation 4.23) relating SBP/SSP to NIV. Panels c) to f) show the posterior parameter estimates obtained from fitting the linear regression model with skew normal residuals estimated using a 12,000 simulation MCMC sample with a burn-in of 4,000 simulations.

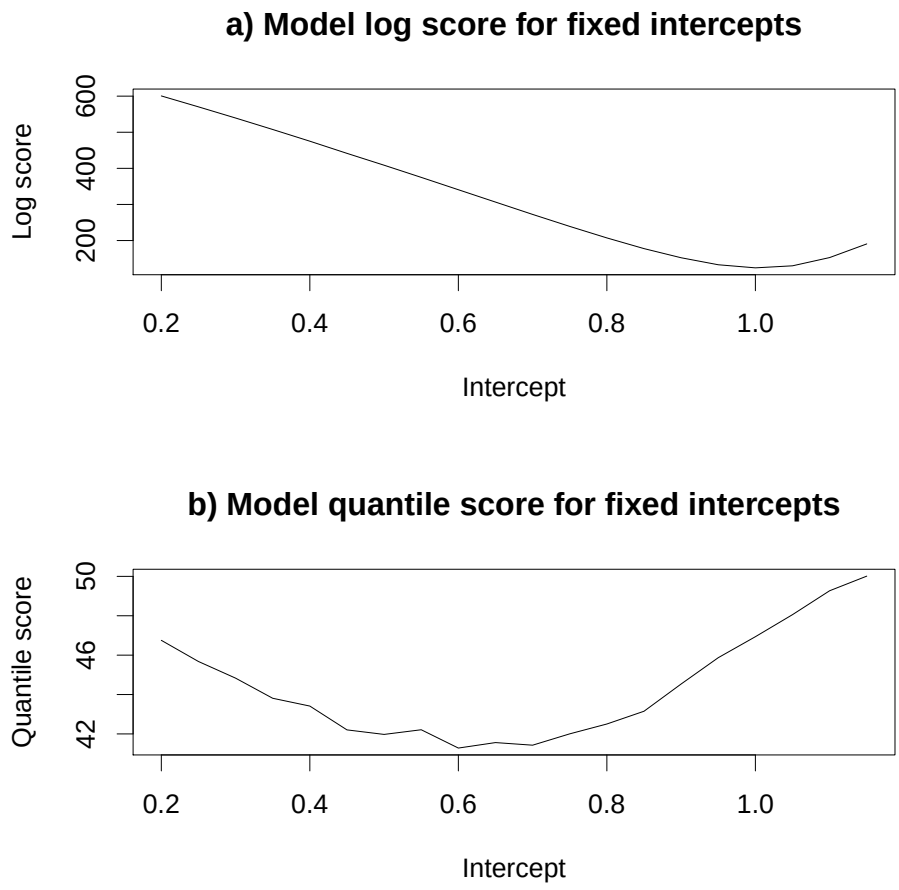

Figure 4.7: Comparison of the BPSIC obtained for models with fixed intercept values in the linear regression. Note that the minimum BPSIC log score is obtained with an intercept value of approximately 1.0, consistent with the posterior estimates in Figure 4.6 when a vague prior is placed on the intercept. However, models fitted with lower values of the intercept have lower BPSIC quantile score, reflecting the increased skewness fitted by these models.

Table 4.3: Comparison of BPSIC for fitted and fixed intercepts - skew-normal model

	Fitted	Intercept = 0.6	Intercept = 1.0
BPSIC(log score)	124.1	340.8	123.9
Asymptotic bias(log score)	3.7	4.4	4.5
BPSIC(percentile score)	46.2	41.5	46.8
Asymptotic bias(percentile score)	2.5	3.4	3.4

4.6 Chapter summary

In this chapter, we introduce a Bayesian score based information criterion which is an estimate of the posterior weighted average of out of sample performance for a user defined scoring function relevant to the problem at hand. The effects of the bias inherent in using the observed data for the dual purposes of posterior updating and model selection are controlled through the incorporation of an asymptotic bias correction which is derived by applying the results in Ando [2007].

By doing this we are able to choose from a range of utility functions and incorporate an appropriate score within the selection procedure. Moreover, we overcome the 'catch up' limitation, discussed in Chapter 2, of using a cumulative score applied directly, as we are assessing the model as updated against the full data sample.

Analysis shows that this criterion can give similar results to cross-validation, but may have advantages, both in terms of requiring less computing time to implement, and also of using the full data sample, rather than requiring observations to be left out. For small data sets and/or large numbers of model parameters this can be an important consideration.

We illustrated an example of the application of the BPSIC to modelling different aspects of the distribution of electricity imbalance prices, and showed that it was possible to provide objective justification for the selection of alternative parameter values where our interest was on fitting to specific quantiles of the distribution.

Chapter 5

Modular model selection

5.1 Introductory remarks

In previous chapters we have considered techniques for the comparison of models where the comparison metric is applied at the level of the overall model output. In an environment where 'big models' are increasingly prevalent, the option of using existing models as components within a larger framework (an 'aggregate model') is attractive. This is because it enables predictive tools to be developed more rapidly, 'tried and tested' models to be recycled, and a more familiar and credible narrative structure to be communicated to the end user.

In Section 1.3.3, we gave an example from energy market modelling, in which models for UK electricity prices are often constructed from component price models of the underlying fuels used for power generation (principally nuclear fuels, coal and gas) together with a model which forecasts demand from industrial and household consumers. A set of more deterministic relationships governing which generation options are preferred in which circumstances then allows the models to be combined into a forecasting tool for the resulting electricity price.

In these situations, we may build a number of larger 'aggregate' models from the available, smaller, component models, and wish to select the best performing. A naive

approach is simply to compare the performance of the models on some future observed data. However, this may ignore richer data available to assess particular components of the model – in the electricity example above, it may be that we have a large amount of data to assess the performance of coal price models, but a more recent history only of nuclear or solar generation.

In these situations the user may require an assessment of how the aggregate model is likely to perform, based on the performance of component models. While scoring the outputs of the aggregate model may be a direct way to proceed in order to tailor an appropriate utility to the analyst's situation (particularly when there are large amounts of training data available) in practice there are often limitations. Data on component performance may have been gathered at different times, and therefore may not permit an holistic model score to be readily generated across the full model.

In addition, the analyst may have good reason to believe that for certain components of the model, training data available is not fully representative of the future regime in which the model will be used and may therefore wish to restrict the performance assessment on a particular component to a subset of the data.

In this chapter we develop an approach which enables performance data on component models, where the joint distribution takes the form of an exponential family, to be combined into an assessment of aggregate model performance. The approach naturally accommodates the analyst's desire to assess component model performance over time periods which best reflect the anticipated future operating environment for that component, and at the same time gives most weight to the outputs of each component which have greatest impact on the overall predictive utility of the aggregate model.

5.2 Exponential family model component selection

5.2.1 Context

For simplicity of exposition we consider the situation where we have an aggregate model which consists of a model for X (parameterised by θ) and an exponential family model

for $Y \mid X$, for which a transformation of X denoted by $\eta(X)$ forms the natural parameter. We also suppose our interest is in the fit of the model on variable Y.

Although this set-up is simpler than networks encountered in practice, it should be noted that the approach we are presenting is concerned with assessing the performance of a component model within a larger model. Assuming that the larger model can be described by means of a series of conditional probability distributions, each represented in exponential family form, then the resulting joint distribution, into which the component model 'feeds' its output, will also be of exponential family form (see, for example, Koller and Friedman [2009]), and hence the results here hold for a wide variety of network structures.

We wish to accommodate both the situation in which the model is constructed as a 'plug in' network in which, for example, the maximum a posteriori (MAP) estimator of X is an input into the model for $Y \mid X$, and also a fully Bayesian approach in which the predictive distribution for X is used within the model for $Y \mid X$. To achieve this we use as a measure for the model fit on Y, a metric based on expected deviance (see, e.g. Spiegelhalter et al. [2002]), where we define the *average deviance* of Y

$$D_{avg}(Y) = E_{f_{true}}\left[E_{f(x|\theta)}\left[-2\log f(y_{obs} \mid x)\right]\right], \tag{5.1}$$

with y_{obs} denoting an actual observation of y under the true data generating process, f_{true}. Where we are considering the 'plug-in' MAP estimator, $\hat{x}$, we interpret the inner expectation in Equation 5.1 as $2\log f(y_{obs} \mid \hat{x})$. For the fully Bayesian model we have:

$$D_{avg}(Y) = E_{f_{true}}\left[\int -2\log f(y_{obs} \mid x) f(x;\theta) dx.\right] \tag{5.2}$$

A naive way to approximate the expectation is to observe the values of y_n directly and compute their predictive density under the modelled distribution

$$D_{avg}(Y) \approx \frac{2}{m}\sum_{j=1}^{m}\int -2\log f(y_j \mid x) f(x;\theta) dx. \tag{5.3}$$

We refer to the approximation obtained using this method as the *direct* score. However, it may be that we have more information (for example, a greater 'track record' of observations) on the performance of our model for X which are effectively ignored if we restrict our evaluation to those where we also observe variable Y. We now consider a methodology in which the average deviance can be built up by scoring the separate component models individually.

5.2.2 Bregman divergences

We represent the probability density of a member of an arbitrary exponential family in the form:

$$f(Y \mid \theta) = \exp(\theta T(Y) - A(\theta) + C(y)), \tag{5.4}$$

and make use of the property relating the expectation of the sufficient statistic, T, under the distribution f to the Jacobian of the log normaliser, A:

$$E_{f(Y|\theta)}\left[T(Y)\right] = \nabla A(\theta). \tag{5.5}$$

We define the Bregman divergence corresponding to a convex and differentiable function F as

$$B_F(x \mid\mid y) = F(x) - F(y) - (x - y).\nabla F(y). \tag{5.6}$$

It is well known (see, for example, Nielsen and Nock [2010]) that the Kullback-Leibler divergence between two members of the same exponential family is given by the Bregman divergence corresponding to the log normalizer between the two parameter values. Here we establish slightly more general conditions for this to hold in the case of the expected log score difference between two members of the same exponential family under an arbitrary true distribution. The conditions require that the expectation of the sufficient statistics are equal to those under the true data generating process, the latter expectation which we denote by $E_{f_{true}}$. This will enable us to apply the result to the situation where we are dealing with a model which is not the 'true model', but where we believe that its sufficient statistics are 'calibrated' to the true data generating process.

Lemma 9 *Suppose two members of the same exponential family, as parameterised in Equation 5.4 above, are denoted by $f(Y \mid \theta_1)$ and $f(Y \mid \theta_2)$. If under the true distribution of Y, we have that $E_{f(Y|\theta_1)}[T(Y)] = E_{f_{true}}[T(Y)]$, then*

$$E_{f_{true}}[\log f(Y \mid \theta_1) - \log f(Y \mid \theta_2)] = B_A(\theta_2 \mid\mid \theta_1). \tag{5.7}$$

Proof.

$$\begin{aligned} E_{f_{true}}[\log f(Y \mid \theta_1) - \log f(Y \mid \theta_2)] &= E_{f_{true}}[T(Y)(\theta_1 - \theta_2) + A(\theta_2) - A(\theta_1)] \quad (5.8) \\ = \quad B_A(\theta_2 \mid\mid \theta_1) &+ E_{f_{true}}[(\theta_2 - \theta_1).(\nabla A(\theta_1) - T(Y))] \end{aligned}$$

Under the assumption that $\nabla A(\theta_1) = E_{f(Y|\theta_1)}[T(Y)] = E_{f_{true}}[T(Y)]$, we have

$$E_{f_{true}}[\log f(Y \mid \theta_1) - \log f(Y \mid \theta_2)] = B_A(\theta_2 \mid\mid \theta_1). \tag{5.9}$$

■

We now use the previous results to establish an alternative way to approximate the average deviance introduced in Section 5.2.1.

Theorem 10 *Suppose*

1. *The distribution $f(X, Y; \theta)$ factorises as $f(Y \mid X)f(X; \theta)$, where $f(Y \mid X)$ takes the exponential family form*

$$f(Y \mid X) = \exp(\eta(X)T(Y) - A(\eta(X)) + C(y)), \tag{5.10}$$

2. *We have a series of observations $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ which we use to score the component model $f(Y \mid X)$,*

3. *We have a a series of (possibly overlapping) observations $x'_1, x'_2, \ldots, x'_m$, which we use to score the component model $f(X)$,*

4. *The model for $Y \mid X$ is well calibrated to the true distribution in the sense that for*

all observations of X, x_{obs}, *we have* $E_{f(Y|x_{obs})}[T(Y \mid x_{obs})] = E_{f_{true}}[T(Y \mid x_{obs})]$.

Then the average deviance of Y *can be approximated as a decomposition of two terms: the achieved log scores on* $f(y \mid x)$ *and posterior averaged Bregman scores on the observations over* $f(X;\theta)$*:*

$$D_{avg}(Y) \approx \frac{2}{m}\sum_{j=1}^{m}\int B_A(\eta(x) \mid\mid \eta(x'_j))f(x;\theta)dx - \frac{2}{n}\sum_{i=1}^{n}\log f(y_i \mid x_i). \tag{5.11}$$

Proof. We express the average deviance as the sum of two terms:

$$\begin{aligned} &E_{f_{true}}\left[\int -2\log f(y_{obs} \mid x)f(x;\theta)dx\right] = E_{f_{true}}\left[-2\log f(y_{obs} \mid x_{obs})\right] \\ &+ E_{f_{true}}\left[\int (-2\log f(y_{obs} \mid x) + 2\log f(y_{obs} \mid x_{obs}))f(x;\theta)dx\right]. \end{aligned} \tag{5.12}$$

We approximate the first term on the right hand side as $-\frac{2}{n}\sum_{i=1}^{n}\log f(y_i \mid x_i)$. Using Lemma 1, the second term can be expressed as

$$E_{f_{true}}\left[\int 2B_A(\eta(x) \mid\mid \eta(x_{obs}))f(x;\theta)dx\right] \tag{5.13}$$

which can be approximated by

$$\frac{2}{m}\sum_{j=1}^{m}\int B_A(\eta(x) \mid\mid \eta(x'_j))f(x;\theta)dx. \tag{5.14}$$

■

We refer to the average deviance obtained using this method as the *component* score.

5.2.3 Simulated Example

In the following example, to compare the scores generated by the component and direct methods, we assume that the true data generating process is given by:

$$\begin{aligned} X &\sim Uniform[-4, 4] \\ Y \mid X &\sim Uniform[X-1, X+1] \end{aligned} \tag{5.15}$$

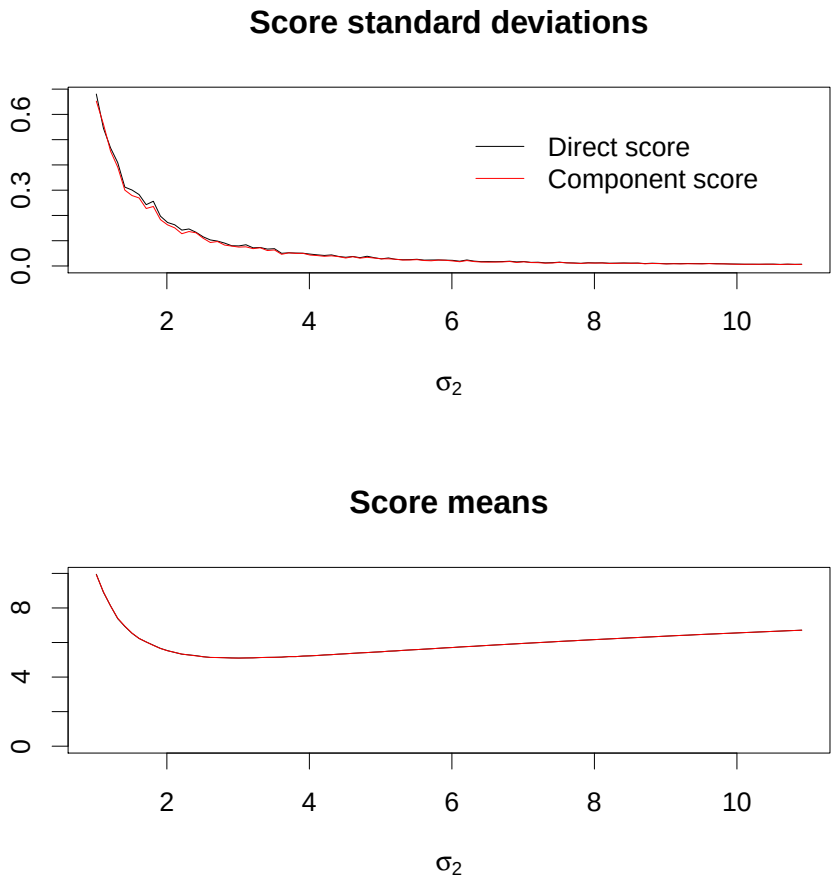

Figure 5.1: Comparison of direct and component scores for a hierarchical normal model for variables X and $Y \mid X$ where the true data generating process is given by a hierarchical uniform model. The figure shows results under different assumptions for the standard deviation σ_2 of the modelled distribution of $Y \mid X$. A sample size of 100 observations of Y and X is used and the means and standard deviations computed across 100 simulations for each value of σ_2.

and that a candidate model has been specified as:

$$X \sim Normal[0, 2] \tag{5.16}$$
$$Y \mid X \sim Normal[X, \sigma_2].$$

Observe that the model for $Y \mid X$ has mean X equal to that of the true distribution for $Y \mid X$, and therefore this satisfies the conditions of Lemma 9. Figure 5.1 compares the direct scores and component scores for a range of values of σ_2 for a sample size of 100 observations of Y and X simulated 100 times for each choice of σ_2. The results are seen to be in close agreement.

The decomposition obtained allows us, when X and Y are both observable quantities, to provide an assessment of the fit of the aggregate model used for predicting Y in terms of two measures: firstly by assessing the component model $f(Y \mid X)$ in terms of its achieved log scores, and secondly assessing the component model $f(X \mid \theta)$ in terms of its achieved average Bregman divergence.

The second of these measures is an expectation of a *Bregman scoring function*. Gneiting [2011] shows that scoring functions of this type correspond to those which are consistent for the mean functional, in the sense that the expected score is minimised when the point forecast corresponding to the mean of the forecast distribution is made.

The decomposition of aggregate model performance on Y immediately provides us with a method to compare candidate component models, f_i, for X for inclusion in the aggregate model: choose the model with the lowest cumulative Bregman score, S_i, which we define as:

$$S_i = \sum \int B_A(\eta(x) \mid\mid \eta(x_{obs})) f_i(x; \theta) dx. \tag{5.17}$$

This approach corresponds to selection of the model for X which is expected to contribute to a lower overall deviance on the aggregate model and therefore provides a model selection metric which is tailored to the properties of the model (as encapsulated in the log normaliser A) to which a component contributes.

5.2.4 Comparison with Bayes Factor selection

It is interesting to compare the approach to standard Bayes factor selection (equivalently the logarithmic score over the joint distribution). To do this, we will assume a more simplified situation where we approximate the posterior averaged integral in Equation 5.17 by a point forecast of X(e.g. the MAP estimate) from $f(x; \theta)$, say $\hat{x}$. We assume that the function η is the identity so that variable X forms the natural parameter in the model for $Y \mid X$. We also assume that the observations for X and $Y \mid X$ coincide (so that the x_i and x_i' terms in Theorem 10 are identical). In this case, our proposed

selection metric can be approximated as choosing the model with the lowest score

$$S_{Comp} = \sum_{i=1}^{n} -\log f(y_i \mid x_i) + B_A(\hat{x} \mid\mid x_i), \tag{5.18}$$

with the Bayes factor (assuming equal prior probabilities) choosing the model with the lowest log score

$$S_{BF} = \sum_{i=1}^{n} -\log f(y_i \mid x_i) - \log f(x_i). \tag{5.19}$$

The term $-\log f(x_i)$ in the Bayes factor score reflects the focus here of selecting of the model for both X and Y which has the highest probability of being the 'true' model; compare this to the term $B_A(\hat{x} \mid\mid x_i)$ in which the focus is on assessing a model for X in terms of its contribution to the model for $Y \mid X$. In general the two approaches will lead to different selection. Within the constraints of the simplifications introduced in this section, the two will correspond only where our model for X has a density function $g(x)$ induced by the divergence B_A:

$$\begin{aligned} g(x) &\propto \exp(-B_A(\hat{x} \mid\mid x)) \\ &\propto \exp(A(x) + (\hat{x} - x)\nabla A(x)). \end{aligned} \tag{5.20}$$

We briefly illustrate different forms of selection metrics and induced densities which would make them equivalent to the Bayes factor selection with reference to three examples of different underlying exponential family models to which the component model provides input.

Example - Normal model with fixed variance σ^2 for $Y \mid X$

In this case, the Bregman score can be written as

$$S_i = \sum \int \frac{(\eta(x) - \eta(x_{obs}))^2}{2\sigma^2} f_i(x;\theta)dx, \tag{5.21}$$

and the selection metric corresponds up to transformation to the log score averaged over the posterior distribution of the parameter θ. The induced density $g(x)$ can be seen by

inspection to be equal to a normal $N(\eta(x_{obs}), \sigma^2)$ density.

Example - Poisson model for $Y \mid X$

In this case, the Bregman score can be written as

$$S_i = \sum \int \eta(x) - \eta(x_{obs}) \left(1 + \log\left(\frac{\eta(x)}{\eta(x_{obs})}\right)\right) f_i(x;\theta)dx, \tag{5.22}$$

and the induced density $g(x) \propto \left(\frac{\eta(\hat{x})}{\eta(x)}\right)^{\eta(x)} e^{\eta(x)}$.

Example - Bernoulli model for $Y \mid X$

In this case, the Bregman score can be written as

$$S_i = \sum \int \log\left(\frac{1-\eta(x_{obs})}{1-\eta(x)}\right) - \eta(x_{obs}) \log\left(\frac{\eta(x)(1-\eta(x_{obs}))}{\eta(x_{obs})(1-\eta(x))}\right) f_i(x;\theta)dx, \tag{5.23}$$

and the induced density $g(x) \propto (1-\eta(x))^{-1} e^{(\eta(\hat{x})-\eta(x))\eta(x)}$.

5.3 Applications

Although in many cases scoring a model directly on the variable of interest may be a sensible way to proceed, we believe there are situations in which the component approach may be advantageous. We describe these in the following sections and illustrate with simulated examples.

5.3.1 Unbalanced data

In many situations there may be more data available to validate one component than the overall model. For example, it may be that there is only recent data where X and Y are observed simultaneously, but a lengthier time period in which X is observed. There is therefore less data on which to assess the model for $Y \mid X$ but more data on which to assess the model for X. By combining the component scores, the use of the full data may allow a reduction in the variance of the estimated score.

To illustrate this we assume that the true data generating process is given by:

$$X \sim Uniform[-2, 2] \tag{5.24}$$
$$Y \mid X \sim Uniform[X - 4, X + 4]$$

and that our model is given by:

$$X \sim Normal[0, 1] \tag{5.25}$$
$$Y \mid X \sim Normal[X, 2]$$

We next examine the impact of having different sample sizes for the number of observations of X, but in each situation, where only the last 20 observations also have associated values of Y. In this case, the direct score therefore uses the sample of 20 observations of Y and X, whereas in the component approach we use this sample to score the model for $Y \mid X$, but use a larger sample, where available, to score the model for X.

Figure 5.2 shows the results of applying the direct and component approaches for different sample sizes for X based on 1,000 simulations for each choice of sample size. As can be seen, as the data available to assess X increases, a reduction in the score standard deviation can be obtained, and therefore the procedure using the component score is likely to be more robust.

In Appendix C we show the results of another simulated example, this time where the model for $Y \mid X$ is given by a Poisson distribution with mean X, and this also demonstrates a reduction in the standard deviation in the score obtained using the component method.

5.3.2 Intervention in an M-partially complete context

In some situations the analyst may have good reason to believe that the data available for model assessment are in some sense unrepresentative of future situations. While it might theoretically be possible to transform these beliefs into a more formal mixture of parametric and non-parametric assumptions (see, e.g. Gutierrez-Pena and Walker [2001],

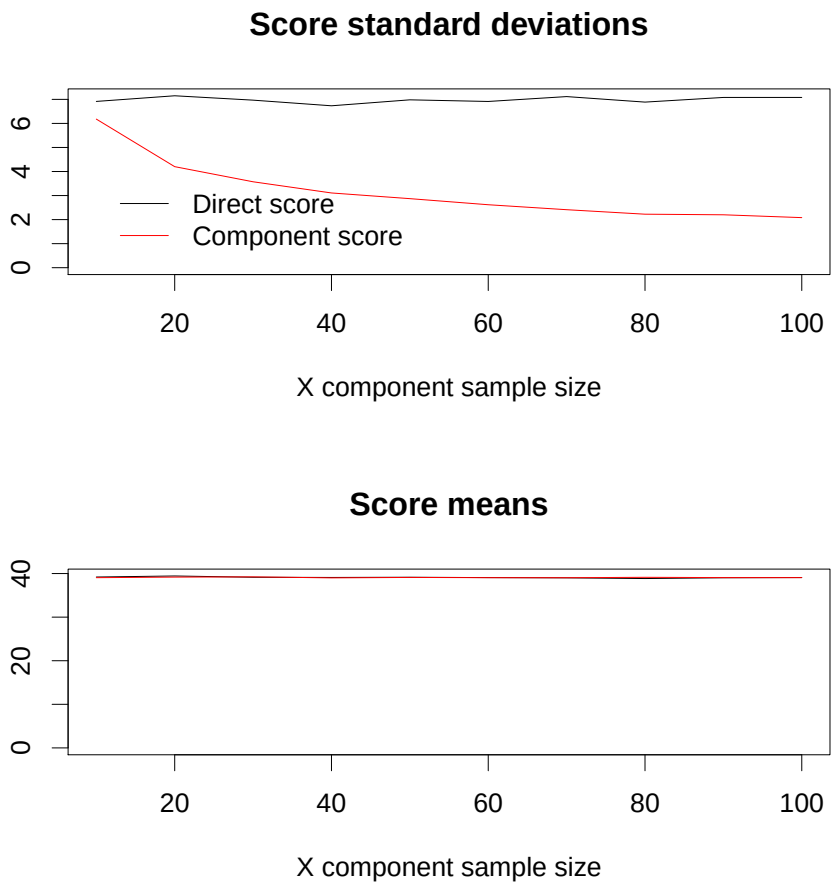

Figure 5.2: Comparison of direct and component scores for the model in section 5.2.3 under differing sample sizes for the X observations, but fixed sample size (20) for the observations of Y.

Gutierrez-Pena and Walker [2005]) and therefore deal with this situation using an M-complete formulation, in practice this is complex. We therefore believe that modular model assessment may provide the analyst with a useful method for intervening within the selection framework in order to make simple modifications to allow additional insights to be incorporated within the final decision.

We refer to the perspective of an analyst who has well-formed beliefs on certain properties of the probability distribution, but is relatively agnostic on others, as an *M-partially complete* perspective, to distinguish it from the M-complete perspective of Bernardo and Smith [1994] in which the analyst has a clear belief model of the full data generating process, but is seeking to choose the best available candidate model as a proxy for her own belief model.

For example, in the electricity markets the price of electricity is set by the highest cost fuel which is used in the generation system. In the UK market the two main sources of generation are coal fired and gas fired generation. Simplifying slightly, at any one time, if coal is higher cost, then the price of electricity will be set at a margin above the price of coal (where the margin reflects the additional plant operating costs); whereas if gas is higher cost, then the price of electricity will be set in relation to the price of gas. Figure 5.3 illustrates this relationship.

Due to the slow mean reverting nature of commodity prices for gas and coal, it can often be the case that one fuel has been observed at a persistently higher cost for a significantly large duration of the data sample period, even though the analyst believes that, for example, a long term forecast would have the probabilities of each fuel being the cost setting fuel at 0.5.

The danger in scoring such a model directly on the observations over this period is that it may disguise poor performance of the model on the the component with a lower contribution to the output. For instance, if the component model for gas is defective, but coal has been the price setting fuel during the observation period, then the model for electricity will score well as the majority of its forecasts are based on the model for coal.

We now show how the indirect approach more naturally accommodates the assessment of the component models for gas and coal in the overall assessment. In equation 5.12 we expressed the average deviance as the sum of two terms:

$$E_{f_{true}}\left[\int -2\log f(y_{obs} \mid x) f(x;\theta)dx\right] = E_{f_{true}}\left[-2\log f(y_{obs} \mid x_{obs})\right] \tag{5.26}$$
$$+ E_{f_{true}}\left[\int (-2\log f(y_{obs} \mid x) + 2\log f(y_{obs} \mid x_{obs}))f(x;\theta)dx\right]$$

In this example, the x variables represent the input fuel (coal or gas) with y denoting the price of power. If we now condition the second term on whether the input fuel is coal or gas (which we denote by x^c and x^g respectively) we have

$$E_{f_{true}}\left[\int 2B_A(\eta(x) \mid\mid \eta(x_{obs}))f(x;\theta)dx\right] \tag{5.27}$$
$$= P(x^c_{obs} > x^g_{obs})E_{f_{true}}\left[\int 2B_A(\eta(x^c) \mid\mid \eta(x^c_{obs}))f(x^c;\theta^c)dx \mid x^c_{obs} > x^g_{obs}\right]$$
$$+ P(x^g_{obs} > x^c_{obs})E_{f_{true}}\left[\int 2B_A(\eta(x^g) \mid\mid \eta(x^g_{obs}))f(x^g;\theta^g)dx \mid x^g_{obs} > x^c_{obs}\right].$$

This now allows the analyst to intervene, by imposing her additional beliefs on the behaviour of gas and power prices on the way in which this expression is approximated. For example, if she believes that in the period over which the model is to be used, we would expect $p(x^g_{obs} > x^c_{obs}) = 0.5$, **and** that there is no difference in model behaviour between the regimes when one fuel is higher priced than the other, then this justifies removal of the conditioning in the expectations, and we can use the approximation:

$$\frac{2}{m}\left[\frac{1}{2}\sum_{j=1}^{m}\int B_A(\eta(x^c) \mid\mid \eta(x^c_{obsj}))f(x^c;\theta^c)dx^c + \frac{1}{2}\sum_{j=1}^{m}\int B_A(\eta(x^g) \mid\mid \eta(x^g_{obsj}))f(x^g;\theta^g)dx^g\right]. \tag{5.28}$$

In this way the full data performance of the component models can justifiably be brought into the computation of the score, even though, perhaps, only one model contributed to the forecasts made during the observation period.

To illustrate these points, we assume that the true data generating processes for gas, coal

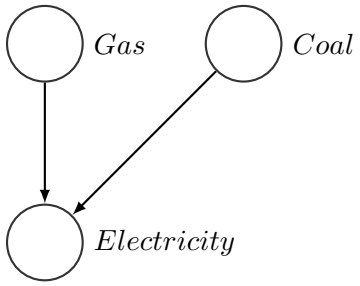

Figure 5.3: Possible representation of a Bayesian network for forecasting the price of electricity

and electricity are given by:

$$\begin{aligned} g_t &= g_{t-1} - \alpha(g_{t-1} - 50) + \epsilon_t^g \\ c_t &= c_{t-1} - \alpha(c_{t-1} - 50) + \epsilon_t^c \\ e_t &= Max(g_t, c_t) + \epsilon_t^p, \end{aligned} \tag{5.29}$$

where $1 \leq t \leq 200$ and the ϵ_t terms are distributed $N(0, 1)$.

Over this period, we simulate the MAP forecasts made by component models for gas and coal and electricity as follows:

$$\begin{aligned} \hat{g}_t &\sim N(g_t, 1) \\ \hat{c}_t &\sim N(c_t, 3) \\ \hat{e}_t &\sim N(Max(g_t, c_t), 2) \end{aligned} \tag{5.30}$$

so that, in particular, we are simulating a situation in which the coal model has a notably higher forecast error than the gas model.

We then simulate the direct and component scores under different mean reversion speeds.

Figure 5.4 shows the results of applying the direct and component approaches for different mean reversion speeds, α, based on simulating 1000 samples for each choice of speed. As can be seen, particularly where the time series are slow to mean revert for low values of α, the standard deviation of the scores computed using the component approach in which we adopt the approximation in Equation 5.28 can be significantly lower. This is because greater use is made of the full set of observed data and forecasts

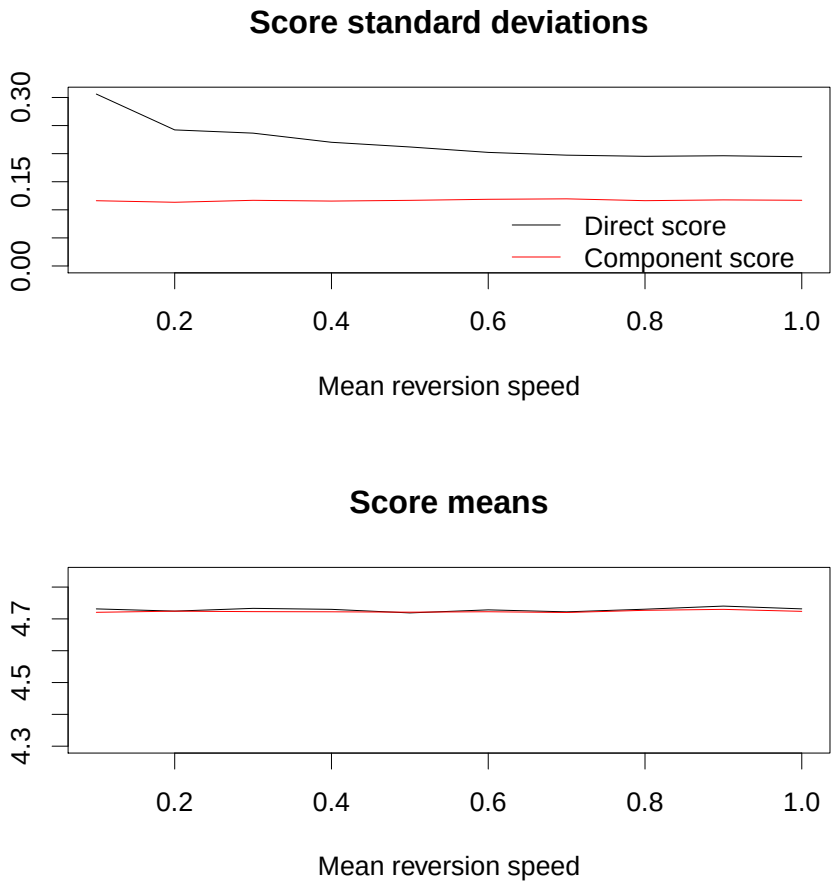

Figure 5.4: Comparison of direct and component model weighted scores for stylised electricity model for different mean reversion speeds for the component gas and coal models.

across all component models.

5.4 Chapter Summary

This chapter has outlined an approach for scoring a model through a combination of component model scores, where the component models were of exponential family form. We believe the approach offers advantages in situations where there may be more data on a particular component model which we wish to include as part of the model assessment. We also showed that it is possible to use such an approach to allow the analyst to 'intervene' in a natural way where she believes that the sample of data is unrepresentative of the future, and has partial beliefs that blur the boundaries between the M-open and M-complete perspectives of Bernardo and Smith [1994].

Chapter 6

Discussion

In this thesis we have argued that, in some contexts encountered in practice, it is appropriate to make modifications to standard model selection criteria. These modifications reflect particular aspects of the context of application.

We have remarked that when utilities linked to forecasting future observations are concerned, and when we do not believe that models under consideration include the true model, the Bayes factor can be sub-optimal, and not reflect the future performance of models, particularly when they have adapted through training on observed data.

In Chapter 3 we have provided examples of situations in which Bayes factor modifications (either through specialising the Bayes factor solely to the variables of interest, or by arranging for likelihoods and priors on 'nuisance' variables to be loosened to reduce their impact on the Bayes factor) may provide improvements when an analyst is interested only in a subset of relationships.

More generally, however, we believe that posterior predictive approaches, approximated using information criteria, have a useful role to play when dealing with general utilities and where computation of cross-validatory alternatives may be prohibitive. Chapter 4 therefore introduced a utility based information criterion. We believe that this is particularly applicable when the analyst has a specific utility reflecting the circumstances in which a model will be used. It may also be useful to employ in situations in which

data is scarce and the analyst wishes to make use of the complete data in model 'training' and selection.

In Chapter 5 we presented initial results on an approach of 'component' scoring, in which a metric of model fit for a composite model was built up from scores on component models. We believe that this may have advantages where component models have been validated over different time periods, or where the analyst wishes to overlay her own views on the relevance of the data to future periods.

For each of these areas, there are a number of related research questions which we would be interested to address in future work:

6.1 Use of 'utility adjusted priors' in model selection

In Chapter 2 we showed that, in some sense, there is a connection between the utility we assume for model selection, and the priors we place within each model. To illustrate this further, we consider the framework in Bernardo and Smith [1994] in which we compare two models M_1, M_2 with the same likelihoods $p(x \mid \theta)$. We assume the models represent complementary hypotheses on the parameter θ, with $M_i : \theta \in \Theta_i$, where $\Theta_1 = \{\theta : \theta \leq \theta_0\}$, $\Theta_2 = \{\theta : \theta > \theta_0\}$ and $\Theta_1 \cup \Theta_2 = \Theta$. The assumption of a single prior $p(\theta)$ can be seen as representing the two priors for individual models, together with the respective probabilities of each model being the true model, in the form of a probability weighted average across models.

If $l_i(\theta)$ denotes a loss function for the parameter θ if model M_i is selected, and we now assume a utility of the form:

$$U(M_i, \theta) = \begin{cases} 0 & \text{if } \theta \in \Theta_i \\ -l_i(\theta) & \text{if } \theta \in \Theta_i^c. \end{cases}$$

then Bernardo and Smith [1994] show that the expected utility of choosing model M_i is equal to:

$$\bar{U}(M_i \mid \mathbf{x}) = -\int_{\Theta_i^c} \frac{l_i(\theta) p(x \mid \theta) p(\theta)}{\int_\Theta p(\mathbf{x} \mid \theta) p(\theta) d\theta}. \tag{6.1}$$

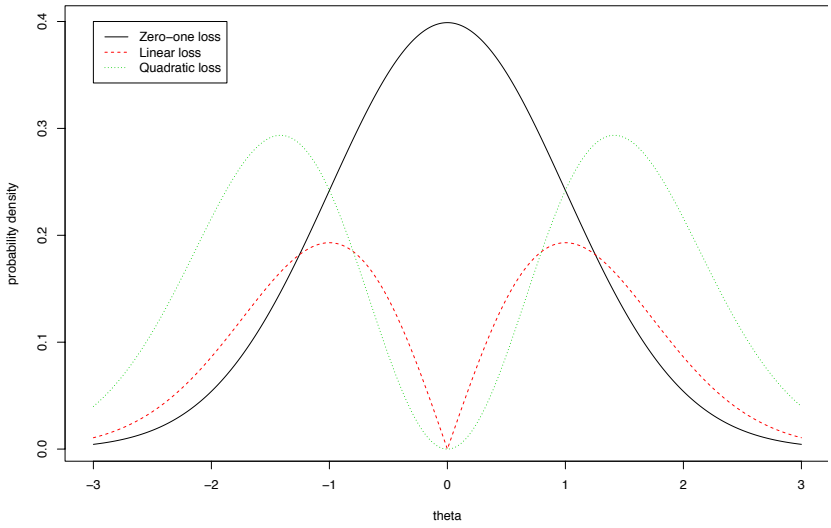

Figure 6.1: 'Pseudo-priors' corresponding to a normal N(0,1) prior. Shown are the zero-one loss, linear loss and quadratic loss.

Ignoring the denominator which is the same for both models, we can see that we prefer model M_1 if

$$\frac{\int_{\Theta_1} p(\mathbf{x} \mid \theta) l_2(\theta) p(\theta) d\theta}{\int_{\Theta_2} p(\mathbf{x} \mid \theta) l_1(\theta) p(\theta) d\theta} > 1. \tag{6.2}$$

For standard Bayes factor selection, we have a zero-one utility so that $l_i(\theta) = 1$ for $\theta \in \Theta_i^c$, and this reduces to the standard 'integrated likelihood' ratio. For other choices of utility, if we define a 'utility adjusted prior'

$$p^\star(\theta) = \begin{cases} \dfrac{l_i(\theta) p(\theta)}{\sum_{i=1}^{2} \int_{\Theta_i^c} l_i(\theta) p(\theta) d\theta} & \text{if } \theta \in \Theta_i^c \end{cases} \tag{6.3}$$

then model selection under the new utility is equivalent to Bayes factor selection with the corresponding utility adjusted prior. Figure 6.1 shows the utility adjusted priors which correspond to different choices of loss function.

We believe further work on this idea might be useful for a number of reasons. Firstly, we may be making use of standard software packages with built in routines for calculating Bayes factors, but with no functionality to calculate restrictions of the Bayes factor. In this case, being able to 'flex' the utility by changing prior may be a useful way to

proceed. Additionally, for the purposes of prior elicitation, it may be helpful to recognise variables on which looser priors are going to be placed, as these require less expert input. This enables utility considerations to be taken into account in advance, focusing time and effort on the parts of the model specification which are pertinent to the model selection.

We would also be interested to see whether the extensive literature on 'robustness to prior' would apply to 'robustness to utility', and therefore might provide assurance to model users of the continued preference for a model in new circumstances, where utilities on decisions might be different. This might also assist in determining whether knowledge of the utility function for model selection would help to streamline and refine the amount and quality of prior elicitation for a given model.

6.2 Score based information criteria

Particularly for small data sets, it seems intuitively plausible that cross-validation will suffer when influential elements are omitted, and that the bias corrected approach may have some advantages here. It would be interesting to understand better the trade-offs in these situations. Stone [1977] establishes asymptotic equivalence of leave-one-out cross validation and AIC. A further area of research would be to investigate whether similar results can be obtained between cross-validated scoring rules and the analogous BPSIC measure.

Comparison of the asymptotic bias and true bias, particularly with small data sets has shown that, in some situations, the true bias may be systematically over-estimated or under-estimated. While we have found the size of this discrepancy to be relatively small in comparison to the underlying cumulative score, it would be useful to understand further whether there are further correction terms which might be introduced to compensate.

We suspect that these may be a consequence of the underpinning Laplace and other first order approximations, though further research is required in this area. We would also be interested in determining any relationships between the extent to which the candidate models are close to the 'true' model and the accuracy of the approximation, and any

implications this might have for the application of the approach in M-open and M-closed situations.

It appears that there may be an interesting connection between the amount of bias in updating and assessing a model using the same data, and the similarity between the metric for model assessment and the target divergence which the update seeks to optimise.

For example, standard Bayesian updating results asymptotically in a posterior which minimises the Kullback-Leibler divergence to the true model (Berk [1966]). This is the same metric which we use to assess the model when we use the logarithmic score. We might, therefore, expect the bias here to be greater than if, for example, we were to score the model with a quantile score.

Bissiri et al. [2013] propose application of alternative updating mechanisms to the Bayes rule in M-open contexts. These take the form of $\pi(\theta \mid y) \propto \exp(l(\theta, y))\pi(\theta)$, for a loss function of interest $l(\theta, y)$, acknowledging that in an M-open context it is not necessarily true that one set of parameters will be optimal under all losses. Under such updating procedures we can derive a similar information criterion to the BPSIC, where the relevant matrices and posterior modes are replaced with their analogues under the alternative loss functions. Here we might expect the bias to be greater if the loss function chosen is comparable to the score function.

We have remarked that our methods may be particularly applicable in big data contexts, for example, where model selection could be tailored to reflect the decision maker's utility more accurately by using a BPSIC based on relevant marginal and conditional logarithmic scores of the variables of interest. Another benefit relates to the lower bias correction term which may apply in these situations.

Typically where cumulative (joint) logarithmic scores are concerned, this will be of the order of the number of parameters in the model, say p. By examining the BPSIC bias correction term in Equation 4.5, it will be seen that the corresponding term is $\mathrm{Tr}(J_n^S(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n))$. So when other, more tailored, scores are used the bias adjustment

will typically be significantly lower. This is because $J_n^S(\hat{\theta}_n)$ will take zero values on those parameters which do not have an impact on the utility under consideration. For example, in the case of a conditional distribution where parameters separate, the bias correction term will equate to the number of parameters involved in the representation of the relevant conditional distribution.

We would also be interested to understand further the application of score based information criteria to derive weighting schemes for use in model averaging. Bayesian model averaging typically proceeds by weighting each model by its posterior probability which, as we have observed, makes most sense when viewed from an M-closed perspective. Clyde and Iversen [2013] suggests an M-open approach where weights are chosen such that the average realised utility across a series of 'leave n out' samples is maximised, however this might share similar computational difficulties to cross-validation.

Implicit in Casanova and Ahrens [2009] is the idea of ensemble weighting, where the weights are linked to some measure of model performance. In an analogous way to which model averaging weights are often approximated by the exponentials of their BIC scores which can be regarded as a proxy for posterior probabilities (see, e.g. Wasserman [2000]), we wonder whether there are are any heuristic methods which would provide methods for averaging models with reference to their achieved BPSIC scores.

6.3 Modular model selection

The results presented in Chapter 5 focus on simple structures within exponential families. We would be keen to extend the range of examples to more realistic models with a greater number of nodes and levels.

It is unclear whether relaxation of the exponential family restriction could be achieved in a tractable way, although it would be interesting to explore the family of divergences which would result. Similarly it would be useful to understand whether it is possible to generalise any of the results to the situation in which we use loss functions other than logarithmic loss to score the performance of the overall model.

Another area of interest to us is the extent of the types of beliefs which can be overlaid by the analyst as part of this framework. We presented an example in which the analyst's beliefs around the frequency of two price regimes allowed a better estimate of the future score of a model to be obtained. Other possibilities might include situations where the analyst wishes to use data from one time period to score one part of a model – reflecting her belief that the future conditions under which the model will be used most resembles this particular period – and another time period to score another part of the same model.

We commented that this type of intervention could be regarded as taking place from what we termed a 'M-partially complete perspective', in which the analyst has a well-defined belief as to one aspect of the data generating mechanism (in this case, that the future probabilities of the gas price setting regime and coal price setting regime occurring were equal), but is relatively agnostic on other details. Rather than adopting non-parametric alternatives (see Gutierrez-Pena and Walker [2001], Gutierrez-Pena and Walker [2005]) we were able to use aspects of the structure to score components of the model in an M-open manner, and use our beliefs to suggest how these should be weighted. We would be interested to explore in more detail how models can be structured to allow this kind of intervention to take place, for example, where the analyst wishes to score certain nodes with an M-closed assumption, and combine these with the scores of others obtained using M-open approaches.

Appendix A

Algorithms used to illustrate Bayes factor robustness

In this appendix we detail the algorithms used to undertake the robustness analysis of the Bayes factor sensitivity to the choice of prior. We assume we are comparing two models M_1, M_2, where the likelihood and prior for model M_i are denoted by $f_i(\theta)$, $\pi(\theta)$ respectively.

A.1 Algorithm 1

First, we wish to investigate the behaviour of the Bayes factor as we allow the prior to vary within the following class of densities around the original choice of prior (which we refer to a the base prior).

$$\Pi_i = \{\pi(\theta) : \frac{\pi(\theta)}{k} \leq \pi_i(\theta) \leq k\pi(\theta)\}. \tag{A.1}$$

This class of densities arises naturally in the elicitation of likely error ranges around a subject matter expert's assessment of probability, or can reflect the range in prior opinion across multiple stakeholders: a choice of $k = 3$ would mean that the probability for any event or interval would be at most three times higher or at least a third of the

quoted probability.

We consider robustness analysis of M_1 only, assuming that the prior for the parameter θ in model M_2 is fixed at its base value $\pi_2(\theta)$. In this case the Bayes factor, $B_{12}(x)$, will be maximised when

$$\bar{B}_{12}(x) = \frac{\sup_{\pi \in \Pi_1} \int f_1(x \mid \theta)\pi_1(\theta)d\theta}{\int f_2(x \mid \theta)\pi_2(\theta)d\theta}. \tag{A.2}$$

The posterior distribution $p_1(\theta)$, corresponding to the base prior, $\pi_1(\theta)$ and data x, is given by

$$p_1(\theta) = \frac{f_1(x \mid \theta)\pi_1(\theta)}{\int f_1(x \mid \theta)\pi_1(\theta)d\theta}. \tag{A.3}$$

If we denote the posterior expectation corresponding to p_1 by E^{p_1}, then we have, for any $\pi \in \Pi_1$:

$$E^{p_1}\left[\frac{\pi(\theta)}{\pi_1(\theta)}\right] = \frac{\int \frac{\pi(\theta)}{\pi_1(\theta)} f_1(x \mid \theta)\pi_1(\theta)d\theta}{\int f_1(x \mid \theta)\pi_1(\theta)d\theta}, \tag{A.4}$$

so that, rearranging, we obtain:

$$\int f_1(x \mid \theta)\pi(\theta)d\theta = E^{p_1}[\frac{\pi(\theta)}{\pi_1(\theta)}] \int f_1(x \mid \theta)\pi_1(\theta)d\theta, \tag{A.5}$$

leading to the following expression for the supremum of the Bayes factor:

$$\bar{B}_{12}(x) = \frac{\sup_{\pi \in \Pi_1} E^{p_1}[\frac{\pi(\theta)}{\pi_1(\theta)}] \int f_1(x \mid \theta)\pi_1(\theta)d\theta}{\int f_2(x \mid \theta)\pi_2(\theta)d\theta}. \tag{A.6}$$

This suggests we choose $\pi \in \Pi_1$ to maximise the posterior expectation $E^{p_1}[\frac{\pi(\theta)}{\pi_1(\theta)}]$. Informally, this can be done by considering areas of equal posterior density, and increasing the prior mass on those areas (subject to an upper limit of multiplying by k) of low prior density while reducing the mass (subject to a lower limit of multiplying by $\frac{1}{k}$) to areas of higher prior density. Adjustments are subject to the constraint that the total mass of the adjusted prior should sum to 1.

We can formalise the algorithm as follows. Suppose we have a sample $\theta_1, \theta_2, \ldots, \theta_N$ from the posterior (for example, the output of a MCMC simulation) and can generate a sample $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N$ from the base prior. We assume that the samples have been

arranged in increasing order. We now split the posterior sample into M consecutive intervals of n elements each (that is, each interval has equal estimated expectation under the posterior density),

$$I_1, I_2, \ldots I_M = (\theta^1, \theta^2], (\theta^2, \theta^3], \ldots, (\theta^M, \theta^{M+1}],$$

where $N = Mn$,

$$\begin{aligned} \theta^1 &= \min(\theta_1, \hat{\theta}_1), \\ \theta^{M+1} &= \max(\theta_N, \hat{\theta}_N), \\ \theta^i &= \theta_{n(i-1)} (1 < i < M+1). \end{aligned} \tag{A.7}$$

We compute estimates of the prior probabilities for each I_i,

$$\hat{\pi}_1(I_i) = \frac{1}{N} \sum_{j=1}^{N} 1_{\hat{\theta}_i \in I_i},$$

and place the intervals in order of their estimated prior probabilities, $I^1, I^2, \ldots, I^M$, where $i < j \Rightarrow \hat{\pi}_1(I^i) \leq \hat{\pi}_1(I^j)$.

We then apply the following algorithm to compute the quantity A, which is used to determine the estimate for the prior density, $\hat{\pi}$ which maximises the Bayes factor.

Algorithm 1 Estimate density in ratio family which maximises posterior expectation

```
a = 0 : b = 0
A = 1 : R = M
while A < B do
  while a ≤ b and A ≤ B do
    a ← a + kπ̂1(I^A)
    A ← A + 1
  end while
  while b > a and A ≤ B do
    b ← b + (1/k)π̂1(I^B)
    B ← B − 1
  end while
end while
return A
```

We estimate the maximising prior $\hat{\pi} \in \Pi_1$ by

$$\hat{\pi}(I^i) = k\hat{\pi}_1(I^i), i \leq A \tag{A.8}$$
$$\hat{\pi}(I^i) = \frac{1}{k}\hat{\pi}_1(I^i), i > A,$$

and the estimated maximum posterior expectation is then given by

$$\hat{E} = \frac{Ak}{M} + \frac{M-a}{kA}. \tag{A.9}$$

A.2 Algorithm 2

In order to investigate robustness of the 'leave one out' Bayes factors considered at the end of Section 2.4.2, the following algorithm is presented.

We consider a more general density ratio class than that considered in the previous algorithm. This is defined by non-negative functions $a(\theta), b(\theta)$, where the density ratio class S consists of the set of prior distributions with kernel densities $p(\theta)$ satisfying

$$a(\theta) \leq p(\theta) \leq b(\theta). \tag{A.10}$$

Recall that the leave one out Bayes factor, $B_{12}(x_j \mid x_{(j)})$ was the Bayes factor for the single observation x_j based on *updated* models incorporating data from the remaining observations $x_{(j)}$.

First observe, as in Algorithm 1, we fix the prior $\pi_2(\theta)$ in model M_2, and consider the impact on the Bayes factor as we allow the prior $\pi_1(\theta)$ in model M_1 to vary in the density ratio class.

The numerator of the Bayes factor, which is the integrated likelihood of model M_1, takes the form:

$$\int f_1(x_j \mid M_1, \theta)\pi(\theta \mid x_{(j)}), \tag{A.11}$$

which is a posterior expectation resulting from the choice of prior, π, in the density ratio class around the base prior π_1.

We can make use of a result obtained by Geweke and Petrella [1998], which establishes a general way of computing bounds on posterior expectations as the prior is allowed to vary within a density ratio class. Suppose we have observed data x, and wish to compute the upper bound of the posterior expectation of a function $g(\theta)$ of the parameter, as the prior is allowed to vary across this density ratio class

$$\overline{E}\left[g(\theta)\right] := sup_{p \in S} E\left[g(\theta) \mid x\right] = sup_{p \in S} \frac{\int_\Theta g(\theta)L(\theta)p(\theta)d\theta}{\int_\Theta L(\theta)p(\theta)d\theta}. \tag{A.12}$$

Geweke and Petrella [1998] consider the case where we have a means of simulating $\theta_m, (m = 1, 2, \ldots, M)$ from the posterior distribution arising from a fixed prior $\tilde{p}(\theta)$, for example using a MCMC sampling from the posterior distribution (see, e.g. Gelfand and Smith [1990], Tierney [1994]). They show that this can be used to obtain a consistent approximation, $\overline{Q}_M$, of $\overline{E}\,[g(\theta)]$ as follows:

Define, for $m = 1, 2, \ldots, M$ and $l = 1, 2, \ldots, M$

$$\begin{aligned} u_m &= a(\theta_m)/\tilde{p}(\theta_m) \\ v_m &= b(\theta_m)/\tilde{p}(\theta_m). \end{aligned} \tag{A.13}$$

1. Sort $g_m = g(\theta_m)$ into nondecreasing order.

2. Define:
$$Q_l = \frac{\sum_{m=1}^{l} g_m u_m + \sum_{m=l+1}^{M} g_m v_m}{\sum_{m=1}^{l} u_m + \sum_{m=l+1}^{M} v_m}. \tag{A.14}$$

3. Find l such that $g_l \leq Q_l \leq g_{l+1}$

4. Set $\overline{Q}_M = Q_l$

Appendix B

Derivation of some limiting values of Bayes factor from first principles

B.1 Derivation of limiting values of the log Bayes factor

We derive the results on the limiting behaviour of the log Bayes factor for the binomial model and multivariate Gaussian model under conjugate updating.

In both case, we assume two models $M_i (i = 0, 1)$ share the *same* likelihood, so we have:

$$\log p_i(\mathbf{x}) = \log \pi_i(\boldsymbol{\theta}) + \log p(\mathbf{x} \mid \boldsymbol{\theta}) - \log p_i(\boldsymbol{\theta} \mid \mathbf{x}),$$

so that the log Bayes factor

$$\log \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \log \frac{\pi_1(\boldsymbol{\theta}) p_0(\boldsymbol{\theta} \mid \mathbf{x})}{\pi_0(\boldsymbol{\theta}) p_1(\boldsymbol{\theta} \mid \mathbf{x})},$$

where $p_i(\mathbf{x} \mid \boldsymbol{\theta})$, $p_i(\boldsymbol{\theta} \mid \mathbf{x})$, $\pi_i(\boldsymbol{\theta})$ denote, respectively, the marginal likelihood, posterior given data $\mathbf{x}$ and prior under model M_i.

We consider asymptotic approximations making use of Stirling's approximation (see, e.g.

Davison [2003]):

$$\log\Gamma(z) = (z - \frac{1}{2})\log z - z + \frac{1}{2}\log(2\pi) + O(\log(z)). \tag{B.1}$$

We also make use of a Taylor series approximation to the log function, when x is large and $h \ll x$:

$$\log(x+h) \approx \log(x) + h/x + O(x^{-2}),$$

and observe that using the identity (see, e.g. Bernstein [2009])

$$\frac{\partial \log|\mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^T, \tag{B.2}$$

allows us to approximate where $\mathbf{X}, \mathbf{H}$ are symmetric matrices

$$\log(\mathbf{X}+\mathbf{H}) \approx \log(\mathbf{X}) + \mathbf{H}\mathbf{X}^{-1} + O(\mathbf{X}^{-2}).$$

B.1.1 Limiting value of the log Bayes factor for the binomial model

Suppose we have a large sample of size N of variable X taking two values: 0,1, which we choose to model with a binomial likelihood, and conjugate beta priors $\pi_i(\theta) \sim Be(\alpha_i, \beta_i)$. Let c denote the number of observations with value 1 in the sample $x = x_1, x_2, \ldots x_N$.

Proposition 11 *Suppose $c/N \to \mu$ for some $0 < \mu < 1$ as $N \to \infty$ in probability. Then we have, in probability*

$$\log\frac{p_1(x)}{p_0(x)} \to \log\frac{B(\alpha_0,\beta_0)}{B(\alpha_1,\beta_1)} + (\alpha_1 - \alpha_0)\log\mu + (\beta_1 - \beta_0)\log(1-\mu), \tag{B.3}$$

where $B(\alpha,\beta)$ denotes the beta function $\dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

In particular, if the two models share the same prior mean, $\hat{\mu}$, but different prior variance, assuming large enough α_i, β_i, say $\alpha_i = K_i\alpha, \beta_i = K_i\beta$ with K_i large, but

$K_i \ll N$, then

$$\log \frac{p_1(x)}{p_0(x)} \to (K_0 - K_1) \log \left(\left(\frac{\hat{\mu}}{\mu} \right)^{\alpha} \left(\frac{1-\hat{\mu}}{1-\mu} \right)^{\beta} \right) - \frac{1}{2} \log \left(\frac{K_0}{K_1} \right). \tag{B.4}$$

Proof. The log Bayes factor, $\log \dfrac{p_1(x)}{p_0(x)}$

$$\begin{aligned} &= \log \frac{\Gamma(\alpha_1+\beta_1)\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma(\alpha_0+\beta_0+N)\Gamma(\alpha_1+c)\Gamma(\beta_1+N-c)}{\Gamma(\alpha_0+\beta_0)\Gamma(\alpha_1)\Gamma(\beta_1)\Gamma(\alpha_1+\beta_1+N)\Gamma(\alpha_0+c)\Gamma(\beta_0+N-c)} \\ &= \log \frac{B(\alpha_0,\beta_0)}{B(\alpha_1,\beta_1)} + \log \frac{\Gamma(\alpha_1+c)\Gamma(\beta_1+N-c)}{\Gamma(\alpha_1+\beta_1+N)} - \log \frac{\Gamma(\alpha_0+c)\Gamma(\beta_0+N-c)}{\Gamma(\alpha_0+\beta_0+N)}. \end{aligned} \tag{B.5}$$

The second and third terms in equation (B.5) have identical forms. Using Stirling's approximation, as z is large, these can be approximated in the form

$$(\alpha_i + \mu N - \frac{1}{2}) \log(\alpha_i + \mu N) + (\beta_i + (1-\mu)N - \frac{1}{2}) \log(\beta_i + (1-\mu)N)$$

$$-(\alpha_i + \beta_i + N - \frac{1}{2}) \log(\alpha_i + \beta_i + N) \tag{B.6}$$

$$\approx (\alpha_i + \mu N - \frac{1}{2})(\frac{\alpha_i}{\mu N} + \log \mu + \log N) + (\beta_i + (1-\mu)N - \frac{1}{2})(\frac{\beta_i}{(1-\mu)N} + \log(1-\mu) + \log N)$$

$$-(\alpha_i + \beta_i + N - \frac{1}{2})(\frac{\alpha_i + \beta_i}{N} + \log N). \tag{B.7}$$

Terms of order $N \log N$, N, $\log N$ are zero or have constant coefficients, and therefore cancel when the difference between the second and third terms in equation (B.5) are taken. The leading term left over is

$$\alpha_i - \frac{1}{2} \log \mu + \alpha_i \log \mu + \beta_i - \frac{1}{2} \log(1-\mu) + \beta_i \log(1-\mu) - (\alpha_i + \beta_i). \tag{B.8}$$

Substituting for the second and third terms in equation (B.5) gives the result. Equation (B.4) results from approximating the term

$$\begin{aligned}
&\log B(K\alpha, K\beta) = \log\Gamma(K\alpha) + \log\Gamma(K\beta) - \log\Gamma(K(\alpha+\beta)) \\
&\approx (K\alpha - \frac{1}{2})\log K\alpha - K\alpha + (K\beta - \frac{1}{2})\log K\beta - K\beta - (K(\alpha+\beta) - \frac{1}{2})\log K(\alpha+\beta) \\
&\approx K(\alpha\log\alpha + \beta\log\beta - (\alpha+\beta)\log(\alpha+\beta)) - \frac{1}{2}\log K.
\end{aligned}$$

■

Multinomial Case

Next we allow the observations x to take $k \geq 2$ discrete values and consider the multinomial likelihoods $p_j(x;\theta)$ with conjugate Dirichlet priors $\pi_j(\theta) = Dir(k, \alpha_j)$ where $\alpha_j = (\alpha_1^j, \alpha_2^j, \ldots \alpha_k^j)$. Let $c = (c_1, c_2, \ldots, c_k)$ denote the number of observations in each category from the sample.

Proposition 12 *Suppose for each $i = 1, \ldots, k$, $c_i/N \to \mu_i$ in probability, for $0 < \mu_i < 1$ as $N \to \infty$ in probability. Then we have, in probability*

$$\log\frac{p_1(x)}{p_0(x)} \to \log\frac{B(\alpha_0)}{B(\alpha_1)} + \sum_{i=1}^{k}(\alpha_i^1 - \alpha_i^0)\log\mu_i, \tag{B.9}$$

where $B(\alpha)$ denotes the multinomial beta function $\dfrac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)}$.

In particular, where the two models share the same prior means, $\hat{\mu}_i$, but different prior variance, assuming large enough α_i^j, say $\alpha_i^j = K_j\alpha_i$, with K_j large, but $K_j \ll N$ we have

$$\log\frac{p_1(x)}{p_0(x)} \to (K_0 - K_1)\log\left(\prod_{i=1}^{k}\left(\frac{\hat{\mu}_i}{\mu_i}\right)^{\alpha_i}\right) - \frac{k-1}{2}\log\left(\frac{K_0}{K_1}\right). \tag{B.10}$$

Proof. We have

$$\log\frac{p_1(x)}{p_0(x)} = \log\frac{B(\alpha_0)}{B(\alpha_1)} + \log B(\alpha_1 + c) - \log B(\alpha_0 + c). \tag{B.11}$$

As in the binomial case, we observe that the second and third terms are in the same form, and each can be replaced by its asymptotic approximation:

$$\begin{aligned}\log B(\alpha_j + c) &= \sum_{i=1}^{k} \log\Gamma(\alpha_i^j + c_i) - \log\Gamma(\sum_{i=1}^{k}(\alpha_i^j + c_i)) \\ &= \sum_{i=1}^{k} \log\Gamma(\alpha_i^j + c_i) - \log\Gamma(\sum_{i=1}^{k}\alpha_i^j + N) \\ &\approx \sum_{i=1}^{k}(\alpha_i^j + \mu_i N - \frac{1}{2})(\frac{\alpha_i^j}{\mu_i N} + \log\mu_i + \log N) - (\sum_{i=1}^{k}\alpha_i^j + N - \frac{1}{2})(\sum_{i=1}^{k}\frac{\alpha_i^j}{N} + \log N).\end{aligned} \tag{B.12}$$

As for the binomial case, we find that the terms of order $N\log N$, N, and $\log N$ are zero or constant, so cancel when the difference in the second and third terms in equation (B.11) is taken. This leaves the leading terms as

$$\sum_{i=1}^{k}(\alpha_i^j - \frac{1}{2})\log\mu_i. \tag{B.13}$$

Substituting the leading terms for the second and third terms in equation (B.11) gives the result. The proof of equation (B.10) follows using the same argument in the previous section for (B.4). ■

B.1.2 Limiting value of the log Bayes factor for the multivariate normal model

We proceed similarly to the previous case. For the d-dimensional multivariate normal model, we have a conjugate normal inverse Wishart prior (see Murphy [2007] for a useful and comprehensive collection of formulae relating to the conjugate analysis of the Gaussian distribution):

$$\begin{aligned}p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{\mu}_0, \kappa, \boldsymbol{\Lambda}, \nu) &= \frac{1}{Z}|\boldsymbol{\Sigma}|^{-((\nu_0+d)/2+1)}\exp(-\frac{\kappa}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1})), \\ Z &= \frac{2^{\nu d/2}\Gamma_d(\nu/2)(2\pi/\kappa)^{d/2}}{|\boldsymbol{\Lambda}|^{\nu/2}}\end{aligned}$$

Proposition 13 *Suppose the true data generating process of multivariate data of dimension d has mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and that we compare two conjugate normal*

inverse Wishart models, where the prior on model M_i is parameterised by $\boldsymbol{\mu}_i, \kappa_i, \Lambda_i, \nu_i$. Assuming a data sample of size n from the true data generating process is observed, then as $n \to \infty$, the difference in log scores tends to a constant limit of

$$\frac{1}{2}(d\log\frac{\kappa_0}{\kappa_1} + \nu_0\log|\boldsymbol{\Lambda}_0| - \nu_1\log|\boldsymbol{\Lambda}_1| + (\nu_1-\nu_0)(\log|\boldsymbol{\Sigma}| + d\log 2)) \tag{B.14}$$

$$+\frac{1}{2}\operatorname{Tr}((\boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_0 + \boldsymbol{D}_1 - \boldsymbol{D}_0)\boldsymbol{\Sigma}^{-1}) + \sum_{i=1}^{d}(\log\Gamma(\frac{\nu_1+1-i}{2}) - \log\Gamma(\frac{\nu_0+1-i}{2})),$$

where $\boldsymbol{D}_i$ is defined as $\kappa_i(\boldsymbol{\mu}-\boldsymbol{\mu}_i)(\boldsymbol{\mu}-\boldsymbol{\mu}_i)^T$.

Proof. We have the difference in log scores:

$$\log p_1(\mathbf{x}) - \log p_0(\mathbf{x}) = (\log\pi_0(\theta) - \log p_0(\theta \mid \mathbf{x})) - (\log\pi_1(\theta) - \log p_1(\theta \mid \mathbf{x})), \tag{B.15}$$

where $p_i(\mathbf{x} \mid \theta)$, $p_i(\theta \mid \mathbf{x})$, $\pi_i(\theta)$ denote the marginal likelihood, posterior given data $\mathbf{x}$ and prior under model M_i. We examine the first term: $(\log\pi_0(\theta) - \log p_0(\theta \mid \mathbf{x}))$ and consider the kernel and proportionality constants in turn. We denote

$$\begin{aligned}\mathbf{S} &= \sum_{i=1}^{N}(\mathbf{x}_i - \bar{x})(\mathbf{x}_i - \bar{x})^T, \\ \mathbf{C} &= \frac{\kappa_0 n}{\kappa_0 + n}(\bar{x} - \boldsymbol{\mu}_0)(\bar{x} - \boldsymbol{\mu}_0)^T.\end{aligned} \tag{B.16}$$

Taking the difference in the logs for the kernel gives

$$\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}^{-1}) + \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Lambda}_0 + \mathbf{S} + \mathbf{C})\boldsymbol{\Sigma}^{-1}. \tag{B.17}$$

$$-\frac{\kappa_0}{2}\operatorname{Tr}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T) - \operatorname{Tr}(\boldsymbol{\Sigma}^{-1}(\frac{\kappa_0 n}{2}\boldsymbol{\mu} - \frac{\kappa_0}{2}\boldsymbol{\mu}_0 - \frac{n}{2}\bar{y})(\boldsymbol{\mu} - \frac{\kappa_0}{\kappa_0+n}\boldsymbol{\mu}_0 - \frac{n}{\kappa_0+n}\bar{y})^T).$$

Cross terms in $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}$ cancel and terms in $\boldsymbol{\mu}_0\boldsymbol{\mu}_0^T$ cancel with the term in $\mathbf{C}$, leaving

$$\frac{n}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\operatorname{Tr}(\mathbf{S})\boldsymbol{\Sigma}^{-1}. \tag{B.18}$$

As this is constant for both $(\log \pi_i(\theta) - \log p_i(\theta \mid \mathbf{x}))$, it will cancel and can be ignored. Taking the difference in the logs for the proportionality constants gives

$$\frac{1}{2}\log\frac{\kappa_0}{\kappa_0+n} - \frac{\nu_0+n}{2}\log|\mathbf{\Lambda}_0+\mathbf{S}+\mathbf{C}| + \frac{\nu_0}{2}\log|\mathbf{\Lambda}_0| + \frac{nd}{2}\log 2 \tag{B.19}$$

$$+\sum_{i=1}^{d}\log\Gamma(\frac{\nu_0+n+1-i}{2}) - \sum_{i=1}^{d}\log\Gamma(\frac{\nu_0+1-i}{2}).$$

We have

$$\log|\mathbf{\Lambda}_0+\mathbf{S}+\mathbf{C}| = \operatorname{Tr}\log(\mathbf{\Lambda}_0+\mathbf{S}+\mathbf{C}), \tag{B.20}$$

and so, as the matrices are symmetric, and $|\mathbf{S}|$ is large, this can be approximated as

$$\begin{aligned}
&\approx \operatorname{Tr}(\log n\mathbf{\Sigma}) + \operatorname{Tr}(\mathbf{\Lambda}_0+\mathbf{C})\mathbf{S}^{-1} &\text{(B.21)}\\
&\approx d\log n + \log|\mathbf{\Sigma}| + \frac{1}{n}\operatorname{Tr}(\mathbf{\Lambda}_0+\mathbf{C})\mathbf{\Sigma}^{-1}\\
&\approx d\log n + \log|\mathbf{\Sigma}| + \frac{1}{n}\operatorname{Tr}(\mathbf{\Lambda}_0+\mathbf{D})\mathbf{\Sigma}^{-1}.
\end{aligned}$$

Using Stirling's approximation

$$\log\Gamma(z) \approx (z-\frac{1}{2})\log z - z + \frac{1}{2}\log(2\pi), \tag{B.22}$$

we can also express

$$\sum_{i=1}^{d}\log\Gamma(\frac{\nu_0+n+1-i}{2})$$

$$\begin{aligned}
&\approx \sum_{i=1}^{d}\frac{\nu_0+n-i}{2}\log\frac{\nu_0+n+1-i}{2} - \frac{\nu_0+n+1-i}{2} + \frac{1}{2}\log 2\pi &\text{(B.23)}\\
&\approx \sum_{i=1}^{d}(\frac{1}{2}\log 2\pi - \frac{\nu_0+n-i}{2}\log 2 - \frac{\nu_0+n+1-i}{2} + 1) + \sum_{i=1}^{d}\frac{\nu_0+n-1}{2}\log n.
\end{aligned}$$

The result follows by considering the difference between this expression and the corresponding expression for $(\log(\pi_1(\theta)) - \log(p_1(\theta \mid x))$. ■

Appendix C

Simulation example - component and direct model performance on Poisson model

C.1 Poisson model

This appendix presents an additional simulation on a Poisson model which was undertaken to provide a further example of the applications of the techniques in Section 5.2.3 to alternative model formulations, in addition to the results for normal distributions which were presented in the main text.

Here, we assume that the true data generating process is given by:

$$\begin{aligned} X &\sim Uniform[4, 20] \\ Y \mid X &\sim Poisson[X], \end{aligned} \tag{C.1}$$

and that our model is given by:

$$\begin{aligned} X &\sim Normal[12, 2] \\ Y \mid X &\sim Poisson[X]. \end{aligned} \tag{C.2}$$

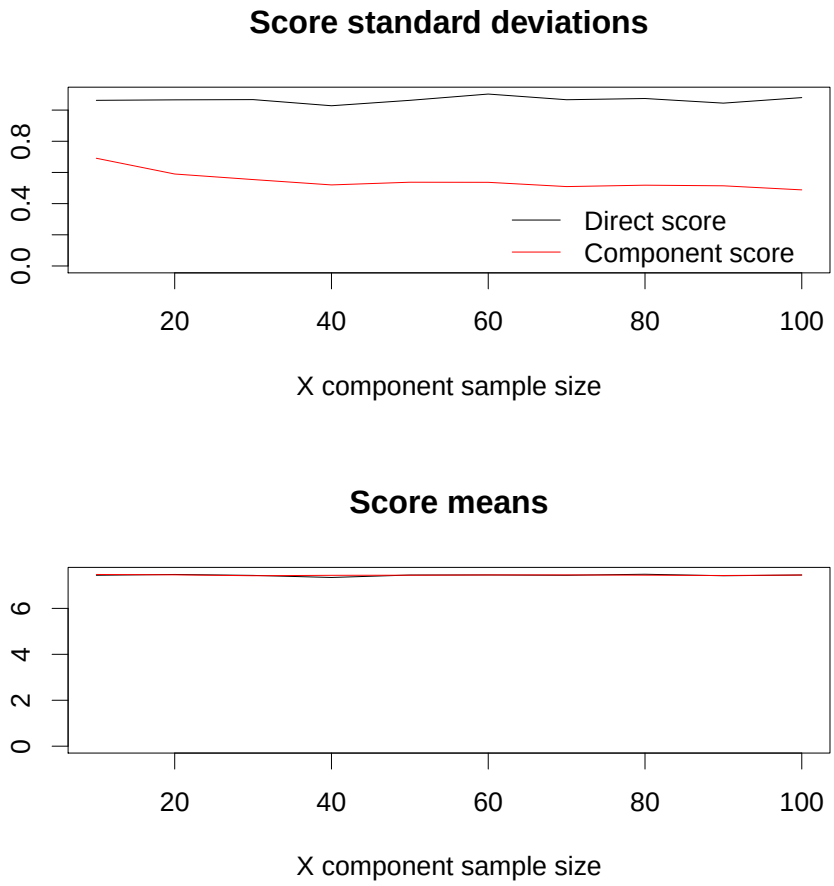

Figure C.1: Comparison of direct and component scores for the model in section 5.2.3 under differing sample sizes for the X observations, but fixed sample size (20) for the observations of Y.

We next examine the impact of having different sample sizes for the number of observations of X, but in each situation, where only the last 20 observations also have associated values of Y. In this case, the direct score therefore uses the sample of 20 observations of Y and X, whereas in the component approach we use this sample to score the model for $Y \mid X$, but use a larger sample, where available, to score the model for X.

Figure C.1 shows the results of applying the direct and component approaches for different sample sizes for X based on 1000 simulations for each choice of sample size. The results support the results in Section 5.3.1 which showed that a reduction in the score standard deviations could be obtained.

References

B. Abramson and A. Finizza. Using belief networks to forecast oil prices. *International Journal of Forecasting*, 7(3):299–315, 1991.

M Aitkin. Posterior Bayes factors. *Journal of the Royal Statistical Society Series B*, 53(1):111–142, 1991.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Second international symposium on information theory*, pages 267–281. Budapest: Akademiai Kiado, 1973.

T. Ando. Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94(2):443–458, 2007.

T. Ando. *Bayesian model selection and statistical modeling.* CRC Press, 2010.

A. Aravkin, A. Kambadur, A. Lozano, and R. Luss. Sparse quantile Huber regression for efficient and robust estimation. *arXiv:1402.4624v1*, 2014.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

F. Asche, P. Osmundsen, and M. Sandsmark. The UK market for natural gas, oil and electricity: are the prices decoupled? *The Energy Journal*, pages 27–40, 2006.

F. Asche, A. Oglend, and P. Osmundsen. Gas versus oil prices: the impact of shale gas. *Energy Policy*, 47:117–124, 2012.

A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12:171–178, 1986.

O.E. Barndorff-Nielsen and D.R. Cox. *Asymptotic techniques for use in statistics.* Chapman and Hall, London, 1989.

F.E. Benth and P.C. Kettler. Dynamic copula models for the spark spread. *Quantitative Finance*, 11(3):407–421, 2011.

J.O. Berger. Robust Bayesian analysis - sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25(3):303–328, 1990.

J.O. Berger and M. Berliner. Robust Bayes and emperical Bayes analysis with ε-contaminated priors. *The Annals of Statistics*, 14:461–486, 1986.

J.O. Berger and L.R. Pericchi. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

J.O. Berger, E. Moreno, L.R. Pericchi, M.J. Bayarri, J.M. Bernardo, J.A Cano, J. De la Horra, J. Martin, D. Rios-Insua, B. Betro, and A. Dasgupta. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.

R.H. Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

J.M Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7:686–690, 1979.

J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, New York, 1994.

D.S. Bernstein. *Matrix mathematics*. Princeton University Press, 2009.

P.G. Bissiri, C.C. Holmes, and S.G. Walker. A general framework for updating belief distributions. *arXiv preprint arXiv:1306.6430*, pages 1–50, 2013.

A. Cartea and T. Williams. UK gas markets: the market price of risk and applications to multiple supply contracts. *Energy Economics*, 30(3):829–846, 2008.

S. Casanova and B. Ahrens. On the weighting of multimodel ensembles in seasonal and short-range weather forecasting. *Monthly Weather Review*, 137(11):3811–3822, 2009.

G. Celeux, F. Forbes, C.P. Robert, and D.M. Titterington. Rejoinder to 'Deviance Information Criteria for Missing Data Models'. *Bayesian Analysis*, 70, 2006.

J.L. Cervera and J. Munoz. Proper scoring rules for fractiles. In *Bayesian Statistics 5*, volume 5. Oxford University Press, 1996.

G. Claeskens and N.L. Hjort. The focused information criterion (with discussion). *Journal of the American Statistical Association*, 98:879–899, 2003.

G. Claeskens and N.L. Hjort. *Model selection and model averaging*. Cambridge University Press, 2010.

M. Clyde and E.S. Iversen. *Bayesian model averaging in the M-open framework*, chapter 24, pages 483–500. Oxford University Press, 2013.

A.C. Davison. *Statistical Models*. Cambridge University Press, 2003.

A.P. Dawid. Statistical theory - the prequential approach. *Journal of the Royal Statistical Society Series A*, 147:278–292, 1984.

A.P. Dawid. The geometry of proper scoring rules. *The Annals of the Institute of Statistical Mathematics*, 59:77–93, 2007.

A.P. Dawid and S.L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable models. *The Annals of Statistics*, 21:1272–1317, 1993.

F. De Santis and F. Spezzaferri. Methods for default and robust Bayesian model comparison: the Fractional Bayes Factor approach. *International Statistical Review*, 67(3):267–286, 1999.

L. DeRobertis and J. Hartigan. Bayesian inference using intervals of measures. *The Annals of Statistics*, 9:235–244, 1981.

D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B*, 57:45–70, 1995.

D. Duffie and J. Pan. An overview of value at risk. *The Journal of Derivatives*, 4 (3):7–49, 1997.

S. Fruhwirth-Schnatter and S. Pyne. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11 (2):317–336, 2010.

S. Geisser. *Predictive inference: an introduction*. Chapman and Hall, 1993.

S. Geisser and W.F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

A.E. Gelfand and D.K. Dey. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society Series B*, 56(3):501–514, 1994.

A.E. Gelfand and S.K. Ghosh. Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85(1):1–11, 1998.

A.E. Gelfand and A.F.M. Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

A. Genz and F. Bretz. *Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, Vol. 195.* Springer-Verlag, 2009.

J. Geweke and L. Petrella. Prior density-ratio class robustness in econometrics. *Journal of Business and Economic Statistics*, 16(4):469–478, 1998.

Z. Gharamani. *Advanced Lectures on Machine Learning*. Springer, 2004.

W.R Gilks, S. Richardson, and D. Spiegelhalter, editors. *Practical Markov Chain Monte Carlo*. New York: Chapman and Hall, 1996.

T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

M. Goldstein. Discusson of "Posterior Bayes Factors" by M.Aitkin. *Journal of the Royal Statistical Society Series B*, 53:134, 1991.

I.J. Good. *Probability and the Weighing of Evidence*. London: Griffin, 1950.

P. Gustafson. Local sensitivity of posterior expectations. *Annals of Statistics*, 24 (1):174–195, 1996.

P. Gustafson and L. Wasserman. Local sensitivity diagnostics for Bayesian inference. *The Annals of Statistics*, 23(6):2153–2167, 1995.

E. Gutierrez-Pena and S.G. Walker. A Bayesian predictive approach to model selection. *Journal of Statistical Planning and Inference*, 93:259–276, 2001.

E. Gutierrez-Pena and S.G. Walker. Statistical decision problems and Bayesian nonparametric methods. *International Statistical Review*, 73:309–330, 2005.

D.J. Hand and V. Vinciotti. Local versus global models for classification problems: fitting models where it matters. *The American Statistician*, 57:124–131, 2003.

B.E. Hansen. Challenges for econometric model selection. *Econometric Theory*, 21: 60–68, 2005.

T. Hastie, R. Tibshirani, and J. Friedman. *Unsupervised Learning*. New York: Springer, 2009.

D. Haughton. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16:342–355, 1988.

D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

S. Howison and M. Coulon. Stochastic behaviour of the electricity bid stack: from fundamental drivers to power prices. *The Journal of Energy Markets*, 27(2):29–69,

2009.

H. Jeffreys. *Theory of Probability*. Oxford University Press, 3rd edition, 1961.

J.B. Kadane and J.M. Dickey. Bayesian decision theory and the simplification of models. In J. Kmenta and J. Ramsey, editors, *Evaluation of Econometric Models*, pages 245–268. Academic Press, 1980.

J.B. Kadane and N.A. Lazar. Methods and criteria for model selection. *Journal of the American Statistical Association*, 99(465):279–290, 2004.

R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

J.T. Key, L.R. Pericchi, and A.F.M. Smith. Bayesian model choice: What and why? In J. Bernardo, editor, *Bayesian Statistics 6*, pages 343–370. Oxford University Press, 1999.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

S.L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

M. Lavine. An approach to robust Bayesian analysis for multidimensional parameter spaces. *Journal of the American Statistical Association*, 86(414): 400–403, 1991.

D.V. Lindley. A statistical paradox. *Biometrika*, 44:187–192, 1957.

D.V. Lindley. Discussion of 'Posterior Bayes Factors' by M.Aitkin. *Journal of the Royal Statistical Society Series B*, 53:130–131, 1991.

H. Linhart and W. Zucchini. *Model Selection*. John Wiley and Sons, 1986.

D.J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. Winbugs – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337, 2000.

A.L. Madsen, F Jensen, U.B. Kjaerulff, and M. Lang. The Hugin tool for probabilistic graphical models. *International Journal on Artificial Intelligence Tools*, 14:507–543, 2005.

K.P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution, 2007. URL `www.cs.ubc.ca/ murphyk/Papers/bayesGauss.pdf`.

K.P. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

M. Musio and A.P. Dawid. Model selection with proper scoring rules. In *Cambridge Statistics Initiative One-Day Meeting*, 2013.

F. Nielsen and R. Nock. Entropies and cross-entropies of exponential families. In *Image Processing (ICIP) 17th IEEE International Conference*, pages 3621–3624. IEEE, 2010.

A. O'Hagan. Fractional Bayes Factors for model comparison (with discussion). *Journal of the Royal Statistical Society Series B*, 57(1):99–138, 1995.

M. Ong, editor. *The Basel Handbook: a guide for financial practitioners*. Risk Books, 2007.

L.D. Phillips. Requisite decision modelling: a case study. *The Journal of the Operational Research Society*, 33:303–311, 1982.

M. Plummer. Penalized loss functions for Bayesian model comparison. *Biostatistics*, 9:523–539, 2008.

A.E. Raftery, D. Madigan, and J.A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92:179–191, 1997.

D. Rios Insua and F. Ruggeri, editors. *Robust Bayesian Analysis*. Springer, 2012.

A. San Martini and F. Spezzaferri. A predictive model selection criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46:296–303, 1984.

G.E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2): 461–464, 1978.

C.-Y. Sin and H. White. Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics*, 71:207–225, 1996.

J.Q. Smith and A. Daneshkhah. On the robustness of Bayesian networks to learning from non-conjugate sampling. *International Journal of Approximate Reasoning*, 51(5):558–572, 2010.

J.Q. Smith and F. Rigat. Isoseparation and robustness in finite parameter Bayesian inference. *Annals of the Institute of Statistical Mathematics*, 64(3), 2012.

D.J. Spiegelhalter, N. Best, B.P. Carlin, and A. Van Der Linde. Bayesian measures of complexity and fit. *Journal of the Royal Statistical Society Series B*, 64(4): 583–639, 2002.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. Van Der Linde. The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society Series B*, 76(3):485–493, 2014.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B*, 36(2):111–147, 1974.

M. Stone. An asymptotic equivalance of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society Series B*, 39(1):44–47, 1977.

L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1762, 1994.

L. Tierney and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

N.T. Underhill and J.Q. Smith. Context-dependent score based Bayesian information criteria. *Bayesian Analysis*, 11(4):1005–1033, 2016.

A. Van Der Linde. A Bayesian view of model complexity. *Statistica Neerlandica*, 66:253–271, 2012.

T. van Erven, P. Grunwald, and S. de Rooij. Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma. *Journal of the Royal Statistical Society Series B*, 74(2):1–37, 2012.

A. Vehtari. *Bayesian model assessment and selection using expected utilities*. PhD thesis, Helsinki University of Technology, 2001.

A. Vehtari and J. Lampinen. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural computation*, 14(10):2439–2468, 2002.

A. Vehtari and J. Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44:92–107, 2000.

R. Weron. *Modeling and forecasting electricity loads and prices: a statistical approach*. John Wiley and Sons, 2007.

S. Wilhelm and B.G. Manjunath. *tmvtnorm: Truncated Multivariate Normal and Student t Distribution. R package version 1.4-6.*, 2012.

R.L. Winkler, J. Munoz, J.M Bernardo, G. Blattenberger, and J.B. Kadane. Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60, 1996.

W.K. Wong. Backtesting value-at-risk based on tail losses. *Journal of Empirical Finance*, 17(3):526–538, 2010.

X. Xu, P. Lu, S.N. MacEachern, and R. Xu. Calibrated Bayes factors for model comparison. Technical Report 855, Ohio State University, 2011.

L. Yu, S. Wang, and K. Lai. Forecasting crude oil prices with an EMD-based neural network ensemble learning paradigm. *Energy Economics*, 30(5):2623–2635, 2008.

S. Zhou. *Bayesian model selection in terms of Kullback-Leibler discrepancy.* PhD thesis, Columbia University, 2011.